5-2001

An Evaluative Argument-Based Investigation of Validity Evidence for the Utah Pre-Algebra Criterion-Referenced Test

Louise Richards Moulding
Utah State University

AN EVALUATIVE ARGUMENT-BASED INVESTIGATION OF VALIDITY EVIDENCE FOR THE UTAH PRE-ALGEBRA CRITERION-REFERENCED TEST

by

Louise Richards Moulding

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Psychology

Approved:

UTAH STATE UNIVERSITY
Logan, Utah

2001

ABSTRACT

An Evaluative Argument-Based Investigation of Validity

Evidence for the Utah Pre-Algebra

Criterion-Referenced Test

by

Louise Richards Moulding, Doctor of Philosophy

Utah State University, 2001

Major Professor: Dr. Karl R. White
Department: Psychology

This study collected evidence to address the assumptions underlying the use of the Utah Core Assessment to Pre-Algebra (UCAP) to (a) measure student achievement in pre-algebra, and (b) assist teachers in making adjustments to instruction. An evaluative argument was defined to guide the collection of evidence. Each of the assumptions in the evaluative argument was addressed using data from a suburban northern Utah school district. To collect the evidence, test content was examined including item match to course objectives, reliability, and subtest intercorrelations. Analyses of correlations of the UCAP with convergent and discriminant measures were completed using student test data ($\underline{N}$ = 1,461), including an examination of both the pattern of correlations and tests of statistical significance. Pre-algebra teachers ($\underline{N}$ = 12) were interviewed to ascertain the degree to which UCAP results were used to make necessary adjustments to instruction.

It was found that the UCAP was technically sound, but measured only 65% of course objectives. Correlation coefficients were analyzed using pattern comparisons and tests of statistical significance. It was found that the pattern of correlation coefficients and the distinction of convergent and discriminant measures supported the UCAP as a measure of mathematics. Teacher interview data revealed that teachers did not make substantive adjustments to the instruction of pre-algebra based on test scores.

Based on these results it was concluded that the underlying assumptions concerning the use of the UCAP were not fully supported. The lack of complete coverage of the pre-algebra course objectives calls into question the ability of the UCAP scores to be used as measures of student achievement, in spite of the technical quality of the test. There was support for the assumption that the UCAP measures mathematics. There was little evidence that teachers use the UCAP score reports to make meaningful and appropriate adjustments to instruction. More evidence is needed to understand the factors that may have led to this lack of use.

The evaluative argument framework defined in this study provides guidance for future research to collect evidence of the validity of decisions based on UCAP scores.

(163 pages)

DEDICATION

The most important words in this document are for my family: Brett, Erin, and Megan. Without your commitment to this endeavor, I would not have been able to complete it. I hope the accomplishment that this represents for our entire family will bring you the joy and satisfaction that it has brought me. I commit my time and love to you now, and I will support you in the fulfillment your dreams.

ACKNOWLEDGMENTS

I would like to thank Dr. Karl White for his guidance, patience, confidence in my abilities, and persistent demand for high-quality work. I would also like to thank the other members of my committee, Drs. Jim Dorward, Byron Burnham, George Julnes, and Kentaro Hayashi, for their valuable feedback and assistance.

I must acknowledge the power of friendship and thank Dr. Nancy Drickey for her unfailing support throughout this process. I relied on her for emotional, psychological, and technical support.

Louise Richards Moulding

CONTENTS

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION

Statement of the Problem

Since the early 1980s there has been an increasing demand for educational reform. In response to the call for reform and educational improvements, state policymakers have increasingly mandated accountability systems that focus on student achievement of knowledge and skill in key content areas. Thus, departments of education in many states have developed criterion-referenced tests (CRTs) to measure student achievement of state-defined standards and objectives. While the purpose of these tests is similar in most states, test scores are used to make a variety of decisions with varying consequences for students. Uses range from student recognition to denial of promotion or graduation.

In 1985 the Utah Statewide Testing Program (USTP) was legislatively mandated for the purpose of measuring student achievement of the Utah core curriculum. Subjects identified for testing were mathematics, language arts, and science at the primary level, with mathematics and science targeted at the secondary level. The test scores were to be used by teachers to make instructional adjustments, with the goal of increasing student achievement. The first edition of the CRTs that comprise the USTP was used in 1987, with a second edition administered beginning in 1997. In 2000 the Utah legislature mandated that districts report USTP test scores as part of a larger school accountability system to be implemented in 2001. Secondary-level mathematics and language arts were identified as subjects of special concern. Because secondary language arts tests have not yet been developed, a secondary mathematics test is the object of this study, specifically

the Utah Core Assessment for Pre-Algebra (UCAP). Pre-algebra is the first mathematics course for secondary age students (7th grade) in most districts, The UCAP, consisting of nine content subtests and three skill subtests, is administered to approximately 30,000 students each year, more students than any other secondary mathematics CRT.

The purpose and use of state tests are associated with underlying assumptions about how well the test measures what it purports to measure, and the extent to which teachers use the scores to make meaningful adjustments to instruction. Collecting evidence to support such assumptions reflects the evaluative argument approach to test validity. Based on a unified concept of validity, the evaluative argument approach identifies the purpose and uses of a test, defines the underlying assumptions, and designs validity studies to collect evidence about the degree to which those assumptions are supported. Thus, deciding whether test scores are valid for a particular purpose is not an "all or none" decision. Instead it is a matter of collecting data until decision makers become sufficiently confident that the test scores are useful for the intended purposes. Therefore, judgments about the validity of purposes and uses are based on accumulation of data from a variety of sources. Investigating the validity of inferences is an ongoing process that must be revisited as the purposes and uses of a test change.

The methods of collecting validity evidence are well established and should be based on the purpose and use of the test. Such methods include, but are not limited to, examination of test item content, correlation of other measures with the criterion-referenced scores, structural equation modeling to explore and confirm theoretical models of variable relationships, multitrait-multimethod research to examine convergent and

discriminant evidence, and experimental research to study the complex relationships involved with instruction, learning, and student achievement as measured by a test.

Even if a test is a good measure of students' knowledge and skills, the test cannot accomplish its intended purpose unless teachers use test scores for making decisions about curriculum and instruction. Collecting evidence about the use of test results to make instructional decisions might include the use of self-reported data from teacher questionnaires, interviews in which teachers describe the interpretation of the test results and the nature of instructional adjustments, and observation of teacher instruction.

Unfortunately, as will be shown in the review of the literature, there is very little evidence to support the purpose and uses of state tests in general or the UCAP in particular. Even less is known about how well teachers use tests to guide instruction, particularly state tests. Therefore, it is unclear whether the UCAP is useful for judging students' mastery of pre-algebra concepts or for assisting teachers in making instructional decisions.

This lack of validity evidence may lead to poor decisions. If the test is a poor measure of mathematics, teachers and administrators may inappropriately promote or retain students, and teachers will lack both understanding of student achievement and the ability to adjust instruction to meet state standards based on the results of the test. If the test is a good measure, but not used by teachers, instruction will not be adjusted and student acquisition of tested knowledge and skill may fall short of state standards.

Therefore, this study was designed to more thoroughly examine the validity of the UCAP for the purposes of (a) measuring seventh graders' mastery of pre-algebra content,

and (b) determining the degree to which UCAP scores are used by teachers to adjust their instruction of the pre-algebra core curriculum.

Purpose and Research Questions

The purpose of this study was to determine the validity and use of the UCAP as a measure of seventh-grade students' knowledge of pre-algebra. Student scores were analyzed to answer two research questions:

1. Is there evidence to support the premise that UCAP scores are an indication of seventh-grade students' mastery of pre-algebra content knowledge and skills?

2. Is there evidence to support the premise that teachers use the results of the UCAP to adjust instruction of the pre-algebra core curriculum?

The first research question was answered through analysis of UCAP test items, measures of internal consistency, and correlation of seventh graders' UCAP scores with other measures. To collect convergent evidence, UCAP scores were correlated with other measures of mathematics achievement: Stanford Achievement Test, 9th Edition (SAT-9) mathematics subtest scores, pre-algebra course grades, and teacher ratings of student mastery of pre-algebra knowledge and skill. Discriminant evidence was provided by correlation of UCAP scores with other SAT-9 subtests including reading, language arts, science, social science, and listening. Conclusions were drawn based on the analysis of correlation patterns and tests of statistical significance.

The second research question was answered through analysis of teacher interview responses to questions concerning (a) receipt of test results, (b) confidence and ability to

interpret results, and (c) use of student scores to inform adjustments to the instruction of the Utah core curriculum for pre-algebra. Data were grouped according to these three areas. Common themes and issues were used to draw conclusions about the use of the UCAP for the purpose of instructional adjustment.

Data for this research were collected from one school district in northern Utah. Thus, to the degree that the students and teachers are different from other students and teachers in Utah, this limitation should be considered when examining the results and discussion of this research.

CHAPTER II

REVIEW OF LITERATURE

The use of tests in schools is a nearly universal aspect of modern education. CRTs in particular have become the focus of teacher and school accountability systems in many states, including Utah. With an increase in the use of tests for making decisions about student achievement, it is imperative that decisions made concerning achievement and instruction are based on test scores that are valid indicators of what was intended to be measured. To the degree that inferences based on these scores are not valid, poor decisions could be made about student achievement or teachers may use scores for making poor instructional decisions. For example, a student may unnecessarily be required to complete a remedial course in pre-algebra prior to advancing to algebra, or a teacher may unnecessarily change an appropriate instructional strategy.

Although the administration of CRTs was mandated in Utah over 10 years ago, little validity evidence had been collected. Prior to conducting this investigation a thorough review of literature was conducted. The review developed a framework for conducting the study by describing the unified concept of validity, developing an evaluative argument framework for collecting validity evidence, reviewing the collection of validity evidence by other states, and reviewing studies that investigated the use of test results by teachers.

Student Tests in State Accountability Programs

Over the past several decades, the effective schools movement has led to accountability systems in most states. State accountability systems often include school accreditation, teacher and administrator evaluation, teacher testing, and student testing. As part of accountability systems, state assessment programs use student achievement testing as a measure of student learning that have become the focus of much attention in the school reform movement.

Calls for Reform

The 1983 release of "A Nation at Risk: The Imperative for Educational Reform" (National Commission on Excellence in Education, 1983) made a call for improvements in the accountability of schools and teachers (Barton, 1999). Many reforms were initiated as a result of studies that followed the release of the document. The reforms "demanded improvement and increased efficiency in the public schools, with the public's concern couched under the broad umbrella of accountability" (Watson, 1990, p. 1).

One consequence of these reforms was the development of and increase in testing programs at the state and school district levels (Barton, 1999; Odden, 1986). Madaus and Tan (1993) also commented on factors responsible for the growth in achievement testing. They contend that three social forces help explain the growth: (a) recurring public dissatisfaction with the quality of education in the United States and efforts to reform education; (b) a broad shift in attention from focusing on resources devoted to education

toward emphasizing results from educational institutions; and (c) an array of legislation, at both federal and state levels, promoting or explicitly mandating standardized testing programs. In particular, states sought to measure achievement of defined core course curriculum and student learning objectives (Council of Chief State School Officers [CCSSO], 1998). Initially, state tests were mandated for use as instructional tools and indicators of educational accomplishments (Baker, 1988; Watson, 1990). However, they are increasingly used to make decisions that have far reaching consequences for students, including course credit, promotion, and graduation. According to Madaus (1987), tests used for important decisions such as these are considered "high-stakes"; "low-stakes" tests are not designed to be central to decision-making, and test performance usually does not result in rewards or sanctions. It is the use of the test scores, not the test, that determines the stakes.

Legislation plays a large role in educational testing. Barton (1999) contends that student testing is the approach of choice for policymakers. Robert Linn in his 1995 Angoff lecture at ETS explained why he believed tests had increased in popularity among policymakers, thus leading to more legislation:

1. Tests and assessments are relatively inexpensive compared to changes that increase instructional time, reduce class size, increase teacher salaries, increase the number of classroom aides, or implementing professional development.

2. Testing and assessment can be externally mandated at the state or district level, which is easier than mandating anything that involves change in what happens inside the classroom.

3. Testing and assessment changes can be rapidly implemented, within the term of elected officials.

4. Results are visible, can be reported to the press, and used to show that legislation led to educational improvements (Barton, 1999, p. 6).

Norm-Referenced and Criterion-Referenced Tests

Two types of tests may be used by states for the purpose of measuring student achievement: norm-referenced tests (NRTs) and criterion-referenced tests (CRTs). Both types of tests can use a number of formats including multiple choice, short answer, extended response, portfolio or projects, and performance assessments. Both types are common in state assessment programs to measure student achievement, and can be used for both high and low stakes decisions. CCSSO (1998) reported that, of the 48 states using student assessments, 31 used NRTs and 33 used CRTs. However, important differences exist in the purpose and use of NRTs and CRTs.

Norm-referenced standardized tests are widely used for national comparison and ranking of students on basic achievement of broad content. The assessments commonly serve as summative assessments of elementary, middle, or high school achievement across broad concepts in key subjects such as mathematics, language arts, science, and social studies (CCSSO, 1998). Norm-referenced tests compare student performance to that of a norming sample. The norming sample is comprised of students who are representative of the intended test takers. An NRT not only reports a percent correct score, but often a percentile score, indicating how well the individual performed compared to a national

group of similar students. Some critics of NRTs argue that they focus on low-level, basic knowledge and do not provide specific information about a student's performance based on a set of local (state or district) standards (Bond, 1995).

By contrast, CRTs measure student achievement by comparing performance to well-defined objectives for a particular content (Hambleton & Rogers, 1991). This form of assessment was first termed criterion-referenced in the 1960s (Glaser, 1963; Popham & Husek, 1969). The purpose of CRTs differs from NRTs. Criterion-referenced instruments are typically constructed to "ascertain an individual's status with respect to a well-defined behavioral domain" (Popham, 1978, p. 93), and/or to differentiate between masters and nonmasters of the content area, or for both purposes. In 1963 Glaser provided the rationale for the use of CRTs:

> Underlying the concept of achievement measurement is the notion of a continuum of knowledge acquisition ranging from no proficiency at all to perfect performance. An individual's achievement level falls at some point on this continuum as indicated by the behaviors he displays during testing. The degree to which his achievement resembles desired performance at any specified level is assessed by criterion-referenced measures of achievement or proficiency. The standard against which a student's performance is compared when measured in this manner is the behavior which defines each point along the achievement continuum. (p. 519)

CRTs have gained wide acceptance since Glaser first coined the term in 1963. Applications of criterion-referenced testing have been used in the classroom, as statewide assessments, as school promotion examinations, and for professional licensure and certification examinations (Hambleton, 1981). In schools, CRTs are most commonly advocated for use in (a) determining student mastery of a set of defined objectives, (b) informing teachers for future instruction of the tested domain, and/or (c) making decisions

about school progress or achievement for the purpose of accountability (Glaser & Nitko, 1971; Haertel, 1985; Hambleton & Rogers, 1991; Hambleton, Swaminathan, Algina, & Coulson, 1978).

In 1998 the CCSSO reported that 32 states were using CRTs to measure mathematics achievement at the secondary level (see Appendix A). This widespread use suggests that legislatures and/or school administrators believe CRTs provide useful information for making decisions about the achievement of mathematics knowledge and skills as defined in state content standards.

Purpose and Use of State Tests

Haertel (1999), in his presidential address at the 1999 annual meeting of the National Council on Measurement in Education, commented on the purposes of testing. He contends that there are three purposes for state testing: (a) to provide information for accountability, evaluation, or comparative purposes; (b) to focus public and media attention on educational concerns; and (c) to change educational practice by influencing curriculum and instruction or by spurring greater effort on the part of school administrators, teacher, and students. State tests, as part of accountability systems, are indeed used for the first purpose stated by Haertel: to provide information about student achievement for accountability; it is the use of the test scores that varies among states. The most frequently cited use of state test scores is instructional decision making as referred to by Haertel's third stated purpose. Teachers are expected to interpret test scores and make use of the information, whether for the design of remedial work for individual students or

adjustments to subsequent instruction leading to greater student achievement of state-defined content standards (CCSSO, 1998; Nolen, Haladyna, & Haas, 1992; Wilson & Corbett, 1991; Yakimowski, 1996). This use, while not considered high stakes, is at the cornerstone of state accountability systems. It is the belief of policymakers and legislators who mandate such tests that teacher use of test scores will lead to better instruction, improved learning, and thus higher student achievement (Barton, 1999; Black & Wiliam, 1998; Haertel, 1999). This assumes, however, that instruction has a direct and large causal effect on achievement as measured by tests, an assumption that has been repeatedly challenged (Berk, 1988; Haertel, 1985, 1986; Haladyna, Haas, & Nolen, 1989).

Student Recognition

Measuring student achievement is the central purpose of state tests. However, few states simply recognize the achievement; most use the results to assign students to a proficiency level, or to make decisions about the students' ability to succeed in future courses or to qualify for graduation. Recognition of student achievement is a logical use of test scores. Students earn recognition, but teachers, schools, and districts are not held accountable for the performance of students. For example, the California Golden State Exams (GSE), taken voluntarily, recognize students who achieve high honors, honors, and recognition levels of achievement on each examination in a number of subject areas, including mathematics (California Department of Education [CDE], 2000). Students who meet these levels are recognized as Golden State Scholars. All Golden State Scholars receive academic excellence awards from the state, and high honors and honors designees

receive a gold insignia on their diplomas. Notice of success on the GSE becomes part of a student's permanent transcript, signifying high achievement to colleges, universities, and employers (CDE, 2000).

Student Grades, Promotion, and Graduation

An increase in stakes for students is the inclusion of test scores as part of a course grade. The use of state tests as a part of student grades is mandated or allowed by some states. For example, the North Carolina State Board of Education has mandated that a student's end-of-course test score count as part of his or her overall grade; however, the amount of the course grade influenced by the score is a local decision (North Carolina Department of Public Instruction [NCDPI], 1996b).

The stakes for students are even higher in states in which the tests are used for identification of students in need of remediation, decisions about promotion, and exit-level tests for graduation. Texas, Pennsylvania, and Maryland, for example, have mandated remediation for students who perform below an acceptable level on state exams (Texas Education Agency [TEA], 1999; Wilson & Corbett, 1991). Not only is the student accountable, but the school or school district must provide remedial assistance to the student prior to promotion to the next grade or course. Passing state tests as a requirement of graduation has also become a more common use of state tests. By 1998, twenty-two states required passing either a number of subject specific tests or an exit exam based on state curriculum standards, with seven other states reporting development of such tests (CCSSO, 1998).

Teacher and School Accountability

Some state accountability systems use student test results to hold teachers, schools, and districts accountable for student learning. In these cases the scores are used (a) by the teacher to make decisions about classroom instruction; (b) to inform parents, the school board, and the public of student achievement of academic standards through published results; (c) to evaluate programs, schools, or school districts; and/or (d) to determine rewards or sanctions to teachers and schools.

For example, tests used in North Carolina have several purposes beyond the measure of student achievement. Technical Report No. 1 for the end-of-level tests indicates that the test results provide an

> independent, uniform source of reliable and valid information which enables (a) students to know the extent to which they have mastered expected knowledge and skills and how they compare to others; (b) parents to know if their children are acquiring the knowledge and skills needed to succeed in a highly competitive job market; (c) teachers to know if their students have mastered grade-level or subject knowledge and skills in the curriculum and, if not, what weaknesses need to be addressed; (d) community leaders and lawmakers to know if students in North Carolina schools are improving their performance over time and how the students compare with students from other states or the nation; and (e) citizens to objectively assess their return on investment in the public schools. (NCDPI, 1996a, p. 1)

If North Carolina students do not achieve scores at a satisfactory level, test results are used in developing strategies and plans for assisting those students, at the expense of the local school or district.

Kentucky has received significant attention for its high stakes accountability testing system. The Kentucky Instructional Results Information System (KIRIS), the legislatively mandated assessment component of the Kentucky Education Reform Act Accountability

System of 1990, was developed to "drive curriculum, instruction, and school administration to ensure that all schools meet the goals for the Commonwealth's schools" (Western Michigan University [WMU], 1995, p. 1). Through KIRIS, the Commonwealth (a) provides an annual assessment of the performance of Kentucky students at selected grade levels, (b) holds each school accountable for achieving the reform goals, (c) administers economic rewards and sanctions based on the test data and noncognitive information, and (d) promotes and supports the use of performance assessment as an integral part of classroom instruction. Economic rewards are granted to schools that show improvement over a threshold level and the state must deliver assistance and sanctions to schools that do not reach their threshold level (Kentucky Department of Education [KDE], 1997; WMU, 1995).

Barton (1999) emphasized that the primary purpose of testing is better information for teachers, administrators, policymakers, and the public. The information must present results that aid instruction and lead to higher achievement. Other uses, including "to grade schools, to scold schools, and to judge whether other improvements in the education system are having the desired effect"(p. 8) are not reasonable until evidence has been collected to validate these uses. Barton contends that "the use of such tests for accountability without meeting standard and well-known methods of validation amounts to test malpractice" (p. 9). Therefore, it is vital that validity evidence be collected for each purpose and use of state test scores.

Accountability and Testing in Utah

The Utah State Office of Education (USOE) received a legislative mandate in 1985 to develop and implement tests to (a) provide a final checkpoint on the extent to which individual students have mastered the content of a core course, and (b) assist teachers in evaluating the strengths and weaknesses of instruction of the core curriculum and make necessary adjustments (USOE, 1997). Teachers, curriculum specialists, university content area specialists, and administrators were involved in constructing tests to assess the degree to which students were learning the Utah core curriculum. Primary grade end-of-level (EOL) tests were developed for reading, mathematics, and science; secondary end-of-course (EOC) tests were developed for mathematics and science. In 1987 the first edition of these tests was administered to students statewide. The Evaluation and Assessment division of USOE reported that more than one half million tests were administered in 1996 (USOE, 2000). As the Utah core curricula were updated to reflect national standards, tests were revised. The second edition was released in 1997, and is currently in use (USOE, 1997). The Utah Statewide Testing Program (USTP) includes the criterion-referenced core curriculum tests described here, and the norm-referenced SAT-9.

In recent years the Utah legislature has expressed concern about secondary math and language arts achievement. As a result, the legislature mandated the development of secondary language arts tests, and an evaluation of the current mathematics tests. The Utah legislature passed legislation (HB 177) to bring together existing tests that are part of the USTP and new assessments under a single accountability system: Utah Performance

Assessment System (U-PASS). It includes the currently used criterion-referenced EOL and EOC tests for math, science, and language arts, and the SAT-9, and mandates a new writing performance assessment and basic skills graduation test. In addition to the original purpose of the CRTs as a measure of student achievement, and their use for instructional decision making, U-PASS mandated public reporting of the results, school ratings based on student performance, and public identification of schools not meeting state standards.

The legislature recognized that existing tests may not be appropriate for use in the U-PASS program and funded an evaluation of the mathematics tests. The legislatively mandated evaluation, conducted by WestEd, an educational laboratory serving Utah, concluded that the mathematics tests already in use would be appropriate for use as part of U-PASS (WestEd, 1999).

Although investigating the degree to which the various tests in Utah are measuring what they purport to measure is important, especially in light of the new level of stakes of U-PASS, the Utah Core Assessment for Pre-Algebra (UCAP) was selected for this study for several reasons. First, pre-algebra is the first course taken by most secondary students in Utah, and there was a large data set available for the UCAP. The UCAP is also administered to more students than any other EOC mathematics test, approximately 30,000 students each year at the completion of the pre-algebra course (Institute for Behavioral Research in Creativity [IBRIC], 1999). Because the test is most often administered to seventh-grade students, other measures such as the SAT-9 are administered within several months of the UCAP. Finally, the UCAP was part of the legislatively mandated evaluation conducted by WestEd, and was specifically deemed

appropriate for use in U-PASS for making decisions about student achievement, instruction, and school quality.

In Utah, the primary purpose of the UCAP is to measure student achievement of pre-algebra knowledge. Student performance on the UCAP can be used as part of the course grade and to determine the future mathematics course work for students (USOE, 1997). Based on the UCAP scores, teachers are expected to adjust their instruction to meet the core curriculum standards, thus leading to higher student achievement. The U-PASS accountability system will raise the stakes of the UCAP and the other CRTs by publicly reporting the scores, and using the scores to rate schools. Unfortunately, as shown in a later section of this review, very little evidence has been provided by USOE or the WestEd evaluation of the mathematics tests to support the use of the UCAP for these purposes. Evidence is needed to determine if these uses are valid.

Judging the Validity and Usefulness of State Tests

The key to determining if state achievement tests are fulfilling the purposes for which they were created is to determine whether they are measuring what they were designed to measure, and if teachers use the results to make instructional adjustments. This concept, referred to as validity, was described by Messick (1993), as "an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores" (p. 13). The evolution of the concept of validity has led to a unified concept in which evidence is collected using well-known methods. A discussion of the concept of validity,

the evaluative argument-based design of validity studies, and methods used to collect validity evidence are provided in this section of the review.

The Unified Concept of Validity

Validity has been a central focus in test development and research since the early 1900s. During the first decades of the century, test publishers assumed responsibility for conducting validation studies. The majority of the studies were atheoretical (Geisinger, 1992), focusing on either (a) the relation between performance on a particular test and the criterion of interest, or (b) the degree to which test content matched the content of a target domain. These early conceptions of validity were given the terms predictive or concurrent validity, and content validity, respectively.

In the 1950s the American Psychological Association (APA) convened the Committee on Psychological Tests to "specify what qualities should be investigated before a test is published" (Cronbach & Meehl, 1955, p. 57). As a result of the committee's work, construct validity was included along with content validity and predictive and concurrent validity in the Technical Recommendations for Psychological Tests and Diagnostic Techniques, commonly referred to as the Recommendations (APA, 1954). Construct validity is concerned with the validity of inferences made about unobserved variables (constructs) based on observed variables (Pedhazur & Schmelkin, 1991). In short, the areas of content, predictive, concurrent, and construct validity were used widely to describe the empirical evidence gathered concerning tests and the use of test scores.

Revised and published three times since they were first published in 1954 by the APA, the Recommendations (later versions of the document have been titled Standards for Educational and Psychological Testing) have provided direction for the development of tests and validation studies for many years. Changes to this document reflect the continuing movement of the psychometric community to a more unified conceptualization of validity. Two important points made in later versions of the Recommendations and in the literature about validity are as follows:

1. Test scores, not test content, should be the focus for any validity study, suggesting that item content alone is not adequate for interpreting the score (APA, American Educational Research Association [AERA], & National Council on Measurement in Education [NCME], 1974; Hambleton, 1980; Hambleton & Rogers, 1991; Linn, 1980; Messick, 1975). Unlike other concepts of validity, "content validity gives every appearance of being a fixed property of the test ... rather than being a property of test responses" (Messick, 1975, p. 959). This is the chief distinction and limitation of content validity, according to Messick. Validity is an accumulation of data to support the use of a test for a particular purpose. As the purpose or use of a test changes, the validity of decisions based on the test score may also change regardless of the constancy of the item content.

2. Test publishers should continue to report results of validation studies they conduct as part of the test development process. However, test users should recognize that test validation is an ongoing process, and should also assume the responsibility for supporting their interpretation of the meaning of test results (APA, AERA, & NCME,

1985; Angoff, 1988; Cronbach, 1971). School counselors, teachers, or admissions officers will only be able to judge the validity of interpretations and decisions to the degree that evidence has been collected for a specific use and interpretation. A test that yields valid scores for one purpose may not yield valid scores for another purpose. For example, a test designed to identify the highest achieving students in mathematics may not be very useful in designing instruction for those students below the specified level of achievement.

Cronbach (1971) made it clear that validation of inferences made for an instrument calls for an integration of many types of evidence. The validity types found in textbooks and the literature are not independent alternatives that stand alone, but only convenient subdivisions to describe different aspects of an integrated investigation of inferences based on test scores (Pedhazur & Schmelkin, 1991). The common reference to types of validity has been addressed numerous times, however, Messick (1993) made an important distinction between the need for different kinds of evidence and different types of validity. As noted by Messick (1993):

> One or another of these forms of evidence, or combinations thereof, has in the past been accorded special status as a so called "type of validity." But because all of these forms of evidence fundamentally bear on the valid interpretation and use of scores, it is not a type of validity but the relation between the evidence and the inferences drawn that should determine the validation focus. The varieties of evidence are not alternatives but rather supplements to one another. This is the main reason that validity is now recognized as a unitary concept. (p. 16)

Due to the compartmentalization of the concept of validity into types, a common, but erroneous belief exists that one could merely pick any type of validity and sufficiently determine if test scores can be used to make valid decisions (Linn, 1980). In the context of

test evaluation, Cronbach (1988, 1989) emphasized that construct validation cannot produce definitive conclusions and cannot ever be finished. Shepard (1993) agreed, stating that "while the never-concluding nature of construct validation is a truism, the sense that the task is insurmountable allows practitioners to think that a little bit of evidence of whatever type will suffice" (p. 429).

Messick (1975, 1980, 1993) proposed using integrated construct validation strategies to establish the evidential basis for interpreting test scores. Angoff (1988) contends that "construct validation is a process, not a procedure; and it requires many lines of evidence, not all of them quantitative" (p. 26). Therefore, construct validity is the overarching term for validity, generally representing the "evidential basis of test interpretation" (Messick, 1980, p. 1019). The different "types" of validity--predictive and concurrent, content, and so on--should be considered data collection and data analysis strategies used for testing the conceptual connections between the measurement and the construct (Angoff, 1988; Messick, 1980). Data collection strategies must be determined based on the purpose and intended use of the test scores.

Collecting Validity Evidence: An Evaluative Argument Approach

Deciding whether test scores are valid for a particular purpose requires an accumulation of evidence. This evidence is necessary for one to be convinced that a particular use or inference based on a test score is valid. Messick (1993) described validity as

> a matter of degree, not all or none. Inevitably, then, validity is an

> evolving property and validation is a continuing process. Because evidence is always incomplete, validation is essentially a matter of making the most reasonable case to guide both current use of the test and current research to advance understanding of what the test scores mean. (p. 13)

Validity evidence should be collected with a focus on the purpose, uses, score interpretations, and inferences instead of on a type of validity (Linn, 1980). Evidence is needed to support each purpose and use (Angoff, 1988). To determine appropriate methods of collecting validity evidence, the purpose(s) of a test must be clearly established.

According to Cronbach (1971) there are two uses of tests: (a) describing the test taker, and (b) making decisions about the test taker. The description of the test taker, based on the test score, relies on the soundness of the test content and the extent to which the construct is measured, while the decisions about the test taker are usually made based on the expected future performance of the individual or group. Cronbach later expanded the idea of test use as a framework for validation studies when he described validation as an evaluative argument. In the framework of evaluation, relevant questions are collected, priorities are assigned to potential lines of inquiry, then selection of important questions are based on the questions that will yield the most information. "After weighing these criteria, the evaluator will probably choose a few questions for intensive research, with other questions covered incidentally by inexpensive side-studies, or not at all" (Cronbach, 1989, p. 165).

Cronbach (1988), Messick (1989, 1995), Kane (1992), and Shepard (1993) have all described validation as a process of constructing and evaluating arguments for and

against proposed test interpretations and uses, referred to as an evaluative argument by these authors and an interpretive argument by Kane (1992). Kane explained validation as the evaluation of interpretive argument.

> To validate a test score interpretation is to support the plausibility of the corresponding interpretive argument with appropriate evidence. The argument-based approach to validation adopts the interpretive argument as the framework for collecting and presenting validity evidence and seeks to provide convincing evidence for its inferences and assumptions, especially its most questionable assumptions. One (a) decides on the statements and decisions to be based on the test scores, (b) specifies the inferences and assumption leading from the test scores to these statements and decisions, (c) identifies potential competing interpretations, and (d) seeks evidence supporting the inferences and assumptions in the proposed interpretive argument and refuting potential counter-arguments. (Kane, 1992, p. 527)

A specific example used by Kane (1992) illustrates how an interpretive argument framework helps to focus a validity investigation specifically on intended test use(s)--in Kane's example, use of an algebra placement test to assign college students to either a calculus course or a remedial algebra course. Kane first identified the following assertions of test use: (a) the test measures prerequisite skills in algebra, and (b) the test will indicate appropriate placement for students with low test scores and for students with high test scores. Based on these assertions, Kane then identified the following assumptions:

> Assumption 1: Certain algebraic skills are prerequisites for the calculus course in the sense that these skills are used extensively in the calculus course.
>
> Assumption 2: The content domain of the placement test matches the target domain of algebraic skills used in the calculus course.
>
> Assumption 3: Scores on the test are generalizable across samples of items, scorers, and occasions.

> Assumption 4: There are no sources of systematic error that would bias the interpretation of the test scores as a measure of skill in algebra.
>
> Assumption 5: An appropriate measure of success in the calculus course is available.
>
> Assumption 6: The remedial course is effective in teaching the algebraic skills used in the calculus course.
>
> Assumption 7: Students with a high level of skill in algebra would not substantially improve these skills in the remedial course and therefore would not substantially improve their chances of success in the calculus course. (Kane, 1992, pp. 531-532)

Finally, methods for collecting data to address each assumption are designed and used to support the test and its uses.

An Illustrative Example of Arguments and Methods

Similar to Kane's example (1992), an evaluative argument is outlined in this section to provide a framework for a basic validity study of state tests. In the case of state tests of achievement, the basic purpose is to measure student achievement, and the primary use is to aid teachers in adjusting instruction. The underlying assumptions of this purpose and use are listed, followed by examples of methods that could be used to address the assumptions. Neither the assumptions nor the methods described are exhaustive; instead, the framework provides context by which state test validity evidence is evaluated in a later section of this review.

The first part of the argument involves inferences about students' level of knowledge and skill based on test scores. This part of the argument rests on two main assumptions:

Assumption 1. Test content matches the course objectives.

Assumption 2. Answering test items requires skills included in the subject domain and course curriculum.

The second part of the argument claims that teachers will use the test results to adjust instruction, leading to higher student achievement. There are three main assumptions:

Assumption 3. Students' skill levels depend directly on the content and quality of instruction.

Assumption 4. Results are meaningfully presented to teachers based on the course objectives.

Assumption 5. Teachers can interpret test results and select appropriate instructional methods based on the interpretation.

Table 1 summarizes the argument, assumptions, and evidence-collecting methods used in this example.

Assumption 1: Test content is relevant to and representative of course objectives. Content-related evidence by itself is not sufficient evidence to support inferences based on test scores (Hambleton, 1980; Hambleton & Rogers, 1991; Linn, 1980; Messick, 1975). Shepard (1993) used the term "internal components" to describe the characteristics of test items. Examining internal components includes investigation of reliability, indexes of item difficulty, and item review by experts as vital procedures in determining the soundness of test content.

Reliability estimates can be calculated in a variety of methods, determined by the test purpose and development procedure (Crocker & Algina, 1986). Methods for

Table 1

Evaluative Argument Framework Used to Collect Validity Evidence for State Tests

Argument	Assumptions	Illustrative methods
Tests scores indicate students' level of knowledge and skill.	Assumption 1. Test content is relevant and representative of the course objectives.	Item analysis Reliability estimates Expert opinion Cognitive process analysis
	Assumption 2. Answering test items requires skills included in the subject domain and course curriculum.	Correlational analysis including zero-order, regression, ANOVA, factor analysis, path analysis; Experimental research Contrasting group analysis
Teachers will use the test results to adjust instruction, leading to higher student achievement.	Assumption 3. Students' scores depend on the content and quality of instruction.	Experimental research Path analysis Observation of instruction
	Assumption 4. Results are meaningfully presented to allow interpretation by teachers.	Expert opinion Teacher interview Measure of teacher's knowledge testing
	Assumption 5. Teachers can select and implement appropriate instructional adjustments based on scores.	Experimental research Observation of instruction

determining the reliability when two test administrations are possible include alternate forms and test-retest. These methods are particularly important when multiple forms of a test are available. Methods for determining the reliability when only one test administration is possible include split-half and internal consistency. Internal consistency methods commonly used are the Kuder-Richardson tests and alpha coefficient.

Determining which method is most appropriate is dictated by the intended use of the test scores. The test developer should identify the sources of measurement error that would be most detrimental to useful score interpretation and design a reliability study that permits such errors to occur so that their effects can be assessed (Crocker & Algina, 1986). For most state tests the source of error is in the content sampling or flawed items; therefore, tests of internal consistency such as Kuder-Richardson or alpha coefficient are appropriate.

Evaluation of item content provides important evidence that items are representative of the domain of interest and relevant to the purposes of the test (Messick, 1993). Angoff (1988) contends that a test composed of a limited number of items cannot be thought to be exhaustive of subject matter, the items do not exhaust the universe, nor can they be drawn randomly. Expert opinion of item content dominates the methods used for collecting such evidence. However, Cronbach (1971), Linn (1980), Messick (1989, 1993), and Shepard (1993) have all agreed that this is an imperfect means of judging the item. In addition to the judgement of content specialists, Hambleton et al. (1978) proposed empirical techniques to evaluate the items. One technique devised by Hambleton is the use of a rating scale for each item, to be completed by experts. Hambleton et al. (1978) also suggested that content specialists be asked to match items already written to the defined objectives. By doing so, agreement among content specialists can be tested using methods such as the chi-square test for independence.

Analysis of process can be used to further support the items as a measure of certain cognitive processes. For example, Messick (1993) described the use of "think

alouds" to analyze the processes underlying item or task performance, thereby affording multiple approaches to construct representation. Messick has also included the analysis of response time, and task difficulty as components of process analysis. This type of analysis adds to the information derived from content experts in the earlier phase of test development. An example offered by Messick involves the analysis of systematic errors in mathematics problem solving. Procedural errors or misconceptions of the student are analyzed to determine the difficulty of the item, and the needed instructional intervention that is indicated by selection of each choice on a multiple-choice test. This method would yield particularly useful information from the interpretation of test scores by teachers for development of instructional interventions to address misconceptions. This technique may also be used in addressing the following assumption as well.

Assumption 2: Answering test items requires skills included in the subject domain and course curriculum. Correlation spans a wide array of conceptual methods used to collect evidence that a test measures what it purports to measure. Correlation of test scores to behaviors or performance can often be useful, but there are no generally accepted guidelines for what constitutes adequate evidence of score validity through correlational studies (Crocker & Algina, 1986). The correlational methods used must match the underlying assumptions as identified in the evaluative argument.

Messick (1993) described the need for both convergent and discriminant evidence. This type of evidence signifies that the measure in question is coherently related to other measures of the same construct as well as to other variables that it should relate to on theoretical grounds. Convergent evidence may be obtained through correlation of the test

scores with other supposed measures of the same construct, analysis of variance, factor analysis, structural equation models, and path analysis. Discriminant evidence signifies that the measure is not related to other distinctly different constructs to the same degree as it measures the same construct. For example, math scores may be positively correlated with reading scores, but not as strongly as with other math scores. The lack of correlational evidence of a relationship between the measure in question and other measures of distinctly different constructs is critical for discounting plausible rival hypotheses about the relationship of the constructs (Messick, 1993).

A source of validation evidence that combines convergent and discriminant evidence is the multitrait-multimethod matrix. This method of evaluating the validity of a construct measure examines "the extent to which a measure relates more highly to different methods for assessing the same construct than it does to measures of different constructs assessed by the same method" (Messick, 1993, p. 46). The multitrait-multimethod matrix is a correlation matrix of the different constructs of interest, and the different methods for measuring the constructs. Direct convergent evidence is indicated by the coefficients in which "method 1" and "method 2" are correlated for each construct. These coefficients should be higher than those for the correlations between the heterotrait-heteromethods and for the heterotrait-monomethods (Messick, 1993). Lack of convergence across methods could indicate that one or more methods are introducing variance or else that the methods are not measuring the same constructs. The examination of measurement methods is increasingly important as states increase the variety of test formats including performance assessment and direct assessment of skills such as writing.

Messick has also recommended that analysis of group differences and changes over time be conducted to determine if the skills tested are stable measures of achievement. Investigation of contrasts between experts and novices in a content area is important for state tests in which the purpose is to check student achievement. Students considered to be novices or nonmasters of the content should score distinctly lower than students considered to be experts or masters. In addition, Messick recommended that decisions about improvement in achievement require that tests not only get progressively more difficult over time, but also that items tie the sources of difficulty to the cognitive processes and knowledge structures at successive levels (Messick, 1984). Judgments about improvement of groups of students over time require analysis of group performance, analysis of explanations of group performance (such as item bias), and construction of tests that allow for time analysis. Much of this work should be completed during test construction, with clear explanations included in technical manuals.

As the number of constructs and methods under study increases, it is often difficult to use simple examination of correlation coefficients to describe the relationships. Factor analysis can be used to derive from intercorrelations among items or tests, a limited number of underlying component variables that would account for the observed covariation. This technique can be used to support the use and interpretation of subtests (Stevens, 1996) designed to measure specific curriculum objectives on state tests.

Assumption 3. Students' test scores depend on the content and quality of instruction. Experimental procedures may be the most appropriate means of addressing this assumption. Although true experiments using random assignment of subjects may not

always be feasible in school settings, quantitative research can be used. Gall, Borg, and Gall (1996) described methods such as longitudinal studies and causal comparative studies with pre-post test designs that could shed light on the influence of instruction on test performance. Data collected could be analyzed with extensive use of multivariate statistics, including path analysis and structural equation modeling, to further examine the influence of instruction. Rival explanation for test performance should also be investigated, including teaching to the test, alteration of test administration protocol (such as reading test items to students or allowing more time to finish), and conditions of testing. Nolen et al. (1992) described many sources of score pollution, and call attention to the need for more investigation of influences on student test scores. Nolen et al. encouraged this investigation, not to punish those that make use of such practices, but to better understand the influence of quality instruction on test scores.

Assumption 4. Results are meaningfully presented based on the course objectives. Assumption 5. Teachers can interpret results and select appropriate instructional methods based on the interpretation. Methods to address these two assumptions require the investigation of teacher behaviors, knowledge and skill in interpreting test scores, and pedagogical knowledge and skill. Appropriate methods include the observation of teachers over the course of several years to determine the nature, extent, and effectiveness of instructional adjustments. The complexity of these adjustments, and attributing changes in student achievement to instructional adjustments requires extensive research, including methods described earlier.

Use of the Evaluative Argument Approach

This review of methods used to support the assumptions of state tests does not discuss the many other uses of state tests, such as placement in subsequent courses, remediation, and teacher or school rewards or sanction. However, Haertel (1999) contends that many of these uses are simply means of drawing attention to the importance of a state-defined curriculum and instruction to state standards. In other words, these tactics focus teachers' attention on the primary use of the tests: adjusting instruction to lead to higher achievement. The use of the evaluative argument approach for designing validity studies allows the researcher to expose and examine evidence for the underlying assumptions of test use, particularly when the use includes high stakes decisions.

Evidence of Validity and Use of State Tests

The increase in state testing as part of accountability systems requires that evidence be collected to support the inferences and decisions made based on test scores. This section reviews and compares evidence collected for state mathematics tests to the previously outlined evaluative argument.

Eight tests from six states were included in this review. An initial search of all state department of education web sites yielded validity evidence information for ten states. Additional nformation was available for 8 of these 10 states with documents obtained from testing divisions of departments of education, and searches of the ERIC and Wilson Web databases. Tests were included in this review based on availability of detailed information concerning the collection of validity evidence (i.e., a technical manual), and

match of format and content of the tests to the UCAP. All included tests multiple-choice state tests of secondary mathematics, similar to the UCAP in content (pre-algebra or algebra) and grade level (seventh to ninth grade). Two types of tests are included: EOL and EOC. EOL tests are often administered at the conclusion of "benchmark" years, usually fifth and eighth grade, as students exit primary and middle school, respectively. The purpose of EOL tests is to measure student learning of content standards prior to promotion to the next level of schooling. EOL tests are written to match standards that ideally would have been met by the end of the school level regardless of the course work completed. In contrast, EOC tests are given at the conclusion of individual courses in secondary schools and serve the purpose of determining the achievement of specific knowledge and skills defined by standards for the course. Table 2 displays a summary of tests reviewed.

Purpose and Use of Reviewed State Tests

The state mathematics tests described in this review represent both low stakes and high stakes tests. Low and high stakes designations are based on the definition by Madaus (1987) in which low stakes tests are described as those that are not anticipated to be central to decision-making, and test performance usually does not lead to significant rewards or sanctions; high stakes tests are used for important decisions such as student promotion to the subsequent grade or course, and graduation. As discussed earlier in this review of literature, the uses of state tests vary widely, and have varying levels of consequences for students, teachers, and schools. The determination of level of stakes for

Table 2

Summary of Reviewed State Tests

State	Test type and subject	Purpose and use(s)	Stakes for students
California	EOC--Algebra	Measure student achievement Honors on diploma	Low
North Carolina	EOL--8th Grade Pre-algebra	Measure student achievement Information for instruction	Low
	EOC--Algebra	Measure student achievement Information for instruction	Low
Pennsylvania	EOL--8th Grade Pre-algebra	Measure student achievement Information for instruction	Low
Texas	EOL--8th Grade Pre-algebra	Measure student achievement Retention if standard not met Information for instruction	High
	EOC--Algebra	Measure student achievement Graduation denied if standard not met Information for instruction	High
Utah	EOC--Pre-algebra	Measure student achievement Information for instruction	Low
Virginia	EOC--Algebra	Measure student achievement Promotion to subsequent course and/or graduation denied if standard not met Information for instruction	High

Note. EOL = End of Level; EOC = End of Course

this review was based on student consequences. Low stakes test from California, North Carolina, Pennsylvania, and Utah were designated as such due to few consequences for

students based on test results. Higher stakes tests from Texas and Virginia were designated as such based on student consequences that included denial of promotion to subsequent courses, grades, or graduation when state standards were not met.

Validity Evidence Collected by States

The six states included in this review had collected validity evidence. As established earlier, evidence should be collected based on the purpose and use of tests, not to complete a checklist of validity "types." While there was recognition of the unified concept of validity by some states, each state presented evidence under headings of content, criterion-related, and/or construct validity. A summary of the evidence collected is displayed in Table 3. The focus of this review was the quality and extent of evidence collected to support the purpose of state tests as measures of student achievement and the use of tests as aid for instructional adjustment.

Evidence for Assumption 1: Test Content Matches the Course Objectives

In each state evidence was collected to establish the reliability of test items. Two formulas referred to in this review are the Kuder-Richardson 20 (KR_{20}) and Cronbach's alpha. Both of these formulas are based on the principal of determining the ratio of the sum of the item covariance to the total observed score variance, and yield essentially the same results. Reliability coefficients are sensitive to the number of items contained on the test (Crocker & Algina, 1986), and therefore the reliability coefficients of subtests containing different numbers of items cannot be directly compared. To overcome this

Table 3

Summary of Validity Evidence Collected by States

	Evidence present by state							
	Low stakes					High stakes		
Assumptions and evidence	CA	NC EOL	NC EOC	PA	UT	TX EOL	TX EOC	VA
Assumption 1: Test content matches the course objectives								
Item analysis	✓	✓	✓	✓	✓	✓	✓	✓
Reliability (internal consistency)	✓	✓	✓	✓	✓	✓	✓	✓
Expert opinion	✓	✓	✓	✓	✓	✓	✓	✓
Cognitive process analysis								
Assumption 2: Answering test items requires skills included in the subject domain and course curriculum.								
Correlation with other measure of construct		✓ 1, 2	✓ 1, 2			✓ 2	✓ 2	✓ 1
Analysis of contrasting group performance		✓	✓					
Identification and testing of rival hypotheses for test performance								
Assumption 3: Students' scores depend on the content and quality of instruction								
Experimental tests								
Observation/interview								
Assumption 4: Results are reported meaningfully to allow interpretation by teachers								
Expert opinion								
Measure of teacher knowledge of testing								
Assumption 5: Teachers can select and implement appropriate instructional adjustments based on scores.								
Teacher interview and observation								
Experimental research								

Note. 1 = correlation with other math test; 2 = correlation with course grade.

problem, the Spearman Brown prophecy formula was employed to obtain the adjusted estimate of the reliability coefficient of the state tests as if each had the same number of items (Crocker & Algina, 1986). In this review the test reliability coefficients were adjusted as if each test had 70 items. The Spearman Brown prophecy formula is

$$r_{adjusted} = K\, r_{original} / 1 + (K-1)\, r_{original},$$

where K is the ratio of the number of items to which the test is being adjusted (70 in this case) to the number of items on the original test. Table 4 lists the original and adjusted reliability for each test.

The values of the adjusted reliability coefficient are considered high and support the conclusion of the states that the tests have strong internal consistency. No other reliability measures were reported by any of the six states. Texas did mention that test-retest was not used since students only take the test once (TEA, 1999).

Evidence pertaining to the content of test items, as presented in technical manuals or test guides, relied heavily on the test development process, including the opinion of experts that items matched curriculum objectives. The test development process was extremely similar for all state tests reviewed. In each case, whether an outside contractor or state department of education directed test development, content experts were invited to serve on committees to oversee development of test specifications and items. The committees developed tables of specifications based on state mathematics curriculum. Items were then written by professional item writers, teachers, administrators, university content professors or instructors, and/or state assessment personnel. Items were then

Table 4

Original and Adjusted Reliability Coefficients for State Tests

State test	# of items	Original reliability coefficient	Adjusted reliability coefficient
California Golden State Exam	30	.72	.86
North Carolina 8th Grade EOL Test	80	.92	.91
North Carolina Algebra EOC Test	81	.94	.93
Pennsylvania 8th Grade EOL Test	79	.93	.92
Texas 8th Grade EOL Test	60	.91	.92
Texas Algebra EOC Test	40	.86	.91
Utah Pre-algebra EOC Test	80	.93	.92
Virginia Algebra EOC Test	50	.88	.91

Note. Adjusted coefficient was calculated using the Spearman Brown Prophecy formula, and adjusting each test to 70 items. EOL = End of Level, EOC = End of Course.

reviewed and revised by educators in each state. Field testing of the large pool of items was conducted using representative samples of students throughout each state.

Based on the field test data, items were again reviewed and revised by writing committees. Data used in this review process typically included item difficulty and point biserial correlation of item performance to overall test scores. Based on these data, some items were eliminated from the pool. Other items were revised and included in the final forms of the tests. All states developed multiple forms of the tests, although not all forms are used each year. No state gave details of the number of items originally written, discarded, or revised.

A notable exception to this development process was found in North Carolina where item response theory and equating were used to allow comparison of student scores from grade to grade on the EOL tests. The item pool is used to make new forms of the tests each year. In the other states with low stakes tests, test forms are reused from year to year. Therefore, the item development process in North Carolina was substantial enough to support the use of the EOL test for the stated purpose of measuring growth of student achievement.

Evidence for Assumption 2: Answering Test Items Requires Skills Included in the Subject Domain and Course Curriculum

North Carolina, Texas, and Virginia collected evidence identified as "construct validity" evidence. The evidence relied exclusively on correlations of the test scores with other measures (see Table 5). Texas correlated the test scores of both the EOL and EOC test to course grades. Because EOL tests are administered to all eighth graders regardless of the mathematics courses completed by the student, the correlation to course grade was relatively low (.32). The correlation of the EOC test score to course grade was higher (.64). Due to the high stakes nature of the Texas tests, the evidence presented to support the use of the tests, a single correlation to course grades, which may or may not be a reliable measure of mathematics knowledge and skill, is insufficient. The denial of promotion or graduation based on test scores requires that more evidence be presented.

Virginia correlated the EOC test scores with the SAT-9 math subtest. The SAT-9 is a well-established test that serves as an accepted measure of mathematics (Harcourt

Table 5

Correlation of State Test Scores with External Measures of the Construct

State test	External measure	Correlation coefficient
California Golden State Exam	No evidence collected	--
North Carolina 8th Grade EOL Test	ITBS math scores	.78
	NAEP math scores	.70
North Carolina Algebra EOC Test	EOL test scores from previous year	.73
	Course letter grade	.62
Pennsylvania 8th Grade EOL Test	No evidence collected	--
Texas 8th Grade EOL Test	Course grade (pass/fail)	.32
Texas Algebra EOC Test	Course grade (pass/fail)	.64
Utah Pre-algebra EOC Test	No evidence collected	--
Virginia Algebra EOC Test	SAT-9 math score	.53

Note. EOL = End of Level, EOC = End of Course, ITBS = Iowa Test of Basic Skills, NAEP = National Assessment of Educational Progress, SAT-9 = Stanford Achievement Test, 9th Edition.

Brace Educational Measurement [HBEM], 1997b; Haladyna, Haas, & Allison, 1998). The correlation was moderate (.53). The evidence itself is sound, but insufficient for the high stakes use of test scores in Virginia. As in Texas, Virginia denies promotion and graduation based on test scores. States with high stakes tests need more evidence than a correlation with a single measure, especially course grades that may not be reliable measures of student achievement.

More substantial evidence was collected in North Carolina. Correlation of EOL test scores was calculated for two other measures of mathematics: the Iowa Test of Basic Skills (ITBS) and portions of the National Assessment of Educational Progress (NAEP) exam. The ITBS was administered during the same year as the EOL test. In addition, NCDPI received permission to administer items from the NAEP during the same testing period as the EOL tests. Correlation coefficients for the ITBS and NAEP were .78 and .70, respectively. The relatively high correlation between the EOL test and well-established measures of mathematics is strong evidence that the test measures mathematics knowledge and skill. The EOC test scores were correlated with the previously taken EOL test scores for the same students. North Carolina used the strength of the evidence gathered for the EOL test to conclude that the EOC test scores could be used to make valid inferences about student achievement. The correlation between the EOC test scores and EOL test scores was .73. Course grades were also used as a measure of mathematics achievement and had a more moderate correlation of .62. The use of multiple correlation coefficients from a variety of sources strengthens the evidence collected by North Carolina. This is especially true considering the low stakes of the test for students.

Additional evidence was collected by North Carolina using contrasting groups. Contrasting group studies ask teachers to rate each student in the course according to defined performance standards. Teachers completed this rating prior to the field testing of the tests. Once the tests were scored, an analysis of accuracy of ratings was completed. Mean scores increased with each grouping of performance level, however, there was

substantial range overlap. According to the NCDPI EOL test technical manual (1996a), the contrasting group study was also used to set cut points for the performance standards.

Overall, evidence collected by states to support the assumption that tests measure the constructs of the mathematics course was extremely weak. Given the high stakes nature of tests used in Texas and Virginia, one would expect higher quality and a greater amount of evidence be collected. Surprisingly, this was not the case. The single exception was the evidence collected by North Carolina.

Evidence for Assumptions 3-5: Instruction, Results, and Instructional Adjustments

No evidence was collected by any of the states included in this review concerning Assumptions 3-5. This lack of evidence is alarming considering the importance of teacher use of test information to adjust instruction, leading to higher achievement. A review of literature concerning teacher use of tests in general is presented in the next section.

Summary

The validity evidence presented by the six states followed "validity types." The evidence for validity of test scores relied on the test development process, but most technical manuals lacked sufficient detail to determine if the process was sufficient. North Carolina provided more detail than the others about test development, and also used item response theory instead of classical test theory.

Evidence to support use of tests as measures of student knowledge and skill was also insufficient. Correlation coefficients were provided for most states; however, all

except North Carolina relied on single correlation coefficients to support high stakes tests. California, Pennsylvania, and Utah provided no evidence to support the assumption that mathematics was measured by the tests. Finally, the lack of evidence to support the assumption that teachers are using test results to make instructional decisions was most alarming.

Utah, like the other states with low stakes tests, had no evidence to support the use of the UCAP as a measure of mathematics. The reliance on item review to establish the validity of decisions is insufficient. This is especially true considering the increased stakes UCAP will have with the implementation of U-PASS. The external evaluation conducted by WestEd did not collect any additional evidence to support the use of UCAP, but did provide a second opinion on evidence supporting content of test items. The lack of evidence collected by Utah was the catalyst for this study.

Previous Research about Teacher Use of Test Results

No evidence had been collected by states included in this review to support the assumption that state mandated test results are used by teachers to make instructional decisions. In the absence of such evidence, a review of relevant research was conducted to investigate use of test results for instructional decision making by teachers. The search was conducted using the following databases: ERIC, Wilson Web, Dissertation Abstracts, and PsychLit. Keywords and descriptors used were achievement tests, test use,

instructional effectiveness, state programs, testing programs, educational assessment, and state standards. Twenty studies and one review of literature were found.

To accept the review of literature (Etsey, 1997) as comprehensive and of high quality, four important criteria were applied: (a) the included literature was comprehensive or representative of the subject, (b) outcomes of studies were quantified on a common metric, (c) a discussion of how outcomes covary with study characteristics was included, and (d) the basis for the conclusion(s) was explicit and replicable. The review by Etsey (1997) did not possess these aspects of a quality review of literature and therefore a separate review was conducted. Etsey's review of research included 16 articles about teacher use of standardized tests, all of which were included in the original 20 studies found. Although the review did not meet the criteria listed previously, Etsey made the following conclusion concerning test use: (a) teachers use standardized achievement test results on a limited scale to make educational decisions, with the primary use to confirm or supplement what information they already have about their students; (b) a shift seems to have appeared from the traditional uses of standardized achievement tests results from low stakes to high stakes (Etsey, 1997).

> The 20 studies found were narrowed based on three criteria for inclusion in this review: (a) a primary research focus of determining classroom teacher use of test scores for instructional decision making; (b) report of data indicating extent of teacher use; and (c) a publication date of 1980 or later, representing the decades in which literature suggests that accountability testing surged. Seven of the 20 articles found were included based on these criteria. Most of the 20 studies were eliminated due to the lack of empirical data concerning extent of test use, and instead focused on practical issues related to testing. (p. 2)

Study Characteristics

Of the seven studies, three focused on the use of tests in a single state (Nolen et al., 1992; Marso & Pigge, 1992; McMillan, Myran, & Workman, 1999), three compared the use of tests in two or more states (Green & Williams, 1989; Wilson & Corbett, 1991; Yakimowski, 1996), and one did not specify (Salmon-Cox, 1981). Each of the studies was compared on the following study characteristics: type of test, level of test "stakes," type of instrument used to collect data, sampling method, and response rate. Table 6 summarizes the study characteristics.

Tests referred to in the studies varied by type of test used and level at which it was selected for use (state or district). Two studies referred to use of nationally normed standardized tests; two focused on state-developed criterion-referenced tests; one collected data on use of scores from district-selected standardized tests, but did not identify the tests; and two referred to standardized tests in general.

Test stakes refer to the level of consequences for students based on test scores. In all but two studies, stakes were defined and assigned by the author(s). The definition previously cited by Madaus (1987) was used in four of the five studies that made reference to stakes. In the two studies in which stakes were not an issue addressed by the author(s), a reference to a specific test was not made.

In six of the seven studies, data concerning teacher use of test scores was collected using a survey. The remaining study used teacher interviews. Three studies used random sampling, with one of those using stratified samples. Three other studies selected a random sample of schools or districts that were then invited to participate. Once the

Table 6

Summary of Teacher Use Study Characteristics

Author (year)	Test	Test "stakes"	Type of instrument	Sampling method
Salmon-Cox (1981)	Metropolitan Test, Stanford Achievement Test, California Test of Basic Skill	Low	Interview	Not specified
Green and Williams (1989)	General reference to standardized test	Not specified	Survey	Stratified random--grade level taught
Wilson and Corbett (1991)	Maryland and Pennsylvania state developed CRT	High MD Low PA	Survey	Random
Marso and Pigge (1992)	General reference to standardized test	Low	Survey	Volunteer
Nolen, Haladyna, and Haas (1992)	Iowa Test of Basic Skills	High	Survey	Random
Yakimowski (1996)	District selected standardized tests	Not specified	Survey	Random
McMillan, Myran, and Workman (1999)	Virginia Standards of Learning Tests	High	Survey	Volunteer

Note. CRT = Criterion-referenced test, MD = Maryland, PA = Pennsylvania.

school or district had agreed to participate, administrators were asked to distribute surveys to a specified number of teachers. The single study using an interview for data collection did not specify the method of participant selection.

Subject Characteristics

Subject characteristics examined for each study were participants' school level, participants' position, sample size, and response rate. The level of participants were designated as elementary, secondary, or district. Middle school teachers were considered to be secondary teachers. The position of participants was either teacher or administrator. School principals and district testing administrators were included in the administrator designation.

The sample sizes of the studies had a wide range. Salmon-Cox (1981) interviewed teachers and had the smallest sample size of 65 teachers. The largest sample size was 2,444 teachers in the study conducted by Nolen et al. (1992).

Studies using random samples had response rates of 31 - 81%. The studies using invitation and assignment of participants had response rate ranges of 16 - 96%. The single study using interviews did not indicate if there had been refusal to participate. The sample characteristics are summarized in Table 7.

Measurement Characteristics

Although six of the seven studies used surveys to collect data, the question types varied. While this does not directly impact the study characteristics, data analysis and conversion to a common metric for this review were impacted. Three studies asked teachers to specify use of test scores from a list, including statements about instructional decisions. For these three studies, the percentage of teachers selecting each statement was reported. The other four studies used 5-point Likert scales to allow teachers to rate

Table 7

Summary of Teacher Use Sample Characteristics

Author (year)	School level	Participant position	Sample size	Response rate (%)
Salmon-Cox (1981)	Elementary	Teachers	N = 68	Not specified
Green and Williams (1989)	Elementary Secondary	Teachers	N = 555 WY N = 253 LA	81 WY 54 LA
Wilson and Corbett (1991)	Elementary Secondary	Teachers Administrators	N = 207 MD N = 831 PA	96 MD 55 PA
Marso and Pigge (1992)	Elementary Secondary	Teachers	N = 218	92
Nolen, Haladyna, and Haas (1992)	Elementary Secondary	Teachers Administrators	N = 2444	45
Yakimowski (1996)	District	Administrators	N = 84 CA N = 55 CO N = 104 CT N = 59 IL	41 CA 31 CO 63 CT 33 IL
McMillan, Myran, and Workman (1999)	Elementary Secondary	Teachers	N = 722	16

Note. WY = Wyoming, LA = Louisiana, MD = Maryland, PA = Pennsylvania, CA = California, CO = Colorado, CT = Connecticut, IL = Illinois.

statements about instructional decisions. In all four cases, the value of 1 indicated that instructional changes were made "never," "very rarely," or that test scores had "no influence" on instructional decisions. These studies reported results in one of two ways: a

mean rating for each statement or the percent of respondents selecting each point on the Likert scale.

Survey and interview data included in these studies could not be analyzed using a common metric of effect size or gain scores. Instead, a common metric was developed for the review of the seven studies using a definition that considered both the percent of teachers selecting statements about test use and the mean or Likert scale ratings about test use. Table 8 defines the common metric.

Table 9 summarizes the measurement characteristics for the seven studies, including the structure of the data collection, data analysis, rating of teacher use on the common metric, and quality of the study. Criteria for determining the quality of studies are listed in Appendix B.

Table 8

Criteria Used to Assign Quality to Test Use Studies

	Extent of use rating		
Data collection method	0	1	2
Teacher selection of statements representing test use for instructional decisions	< 25%	26 - 50%	> 50%
Mean Likert rating (1-5) of test use for instructional decisions 1 = "never, very rarely" used or scores had "no influence" on instructional decisions, 5 = "always, nearly always" used or test scores had "substantial influence" on instructional decisions.	< 2.0	2.1 - 3.0	> 3.1

Table 9

Summary of Measurement Characteristics, Common Metric, and Quality of Study

Author (year)	Data source	Data analysis	Use metric	Quality
Salmon-Cox (1981)	Selection of statement	Percent selecting	0	B
Green and Williams (1989)	Selection of statement	Percent selecting	WY Elementary - 0 WY Secondary - 1 LA Elementary - 0 LA Secondary - 0	B
Wilson and Corbett (1991)	Likert rating of statement	Percent selecting each point on scale	MD - 2 PA - 1	A
Marso and Pigge (1992)	Likert rating of statement	Mean of Likert scale for each statement	Elementary - 1 Secondary - 0	B
Nolen, Haladyna, and Haas (1992)	Selection of statement	Percent selecting	Elementary - 1 Secondary - 0	A
Yakimowski (1996)	Likert rating of statement	Mean of Likert scale for each statement	CA - 2 CO - 2 CT - 2 IL - 2	B
McMillan, Myran, and Workman (1999)	Likert rating of statement	Percent selecting each point on scale	Elementary - 2 Secondary - 1	A

Note. CRT = Criterion-referenced test, WY = Wyoming, LA = Louisiana, MD = Maryland, PA = Pennsylvania, CA = California, CO = Colorado, CT = Connecticut, IL = Illinois; A = Good to Excellent, B = Fair to Poor.

Author's Results and Conclusions

The nine studies are described here with more detail than was presented in the

previous tables, including the authors' conclusions. The studies are presented in order of publication year.

Salmon-Cox (1981) based her study on the premise that controversy surrounding standardized testing assumed that information generated by tests was used by teachers. To determine if this was the case, Salmon-Cox interviewed 68 elementary teachers. The elementary teachers involved reported that they depended on their own observations, not results of standardized tests, to make decisions about instruction and student academic needs. Nearly half the teachers reported that test information was a supplement to or confirmation of information they already had about students. Only 20% reported that test information was used to reflect on or guide instruction. Those teachers that did report using test results for instructional decisions did so while "rethinking or shaping large-group curricula or instruction rather than any use tied to individual students" (Salmon-Cox, 1981, p. 633). It was concluded that elementary teachers rarely used test information to mold their instruction or curricular content, in spite of growing use of standardized tests. Such test information was not crucial to the process of teacher decision making.

Throughout the 1980s and 1990s, Green and colleagues conducted a series of studies to investigate various aspects of teachers' attitudes toward and uses of tests (for example see Green, 1992; Green & Stager, 1985, 1986). The study that fit the criteria for this review, quantifying teacher use of test information, was conducted by Green and Williams (1989). Teachers in Wyoming and Louisiana were surveyed about use of standardized test results although no specific test was named. Statements about curriculum evaluation were selected by 22.9% of participating elementary teachers and

29.3% of secondary teachers in Wyoming. In Louisiana a similar percentage of elementary teachers selected statements about curriculum evaluation (21.6%), but virtually no secondary teachers selected such statements. The authors concluded that test use was very low and speculated that attitudes toward testing and training in measurement methods influenced teacher responses.

Wilson and Corbett (1991) compared two states that had developed CRTs to measure student achievement. The two states, Maryland and Pennsylvania, used the tests as measures of competency in reading and mathematics. Maryland administered the test beginning in the ninth grade to determine eligibility for graduation. Pennsylvania administered tests in Grades 3, 5, and 8, and used the test to identify students in need of additional classroom instruction who may not have been identified by other means. Wilson and Corbett found that teachers in Maryland used test results to a greater extent than teachers in Pennsylvania. Forty-nine percent of participating Maryland teachers reported "major changes" to course content and pedagogy due to the test and test results. Only 7% of teachers in Pennsylvania reported such changes. In follow-up interviews with district personnel and teachers in both states, Wilson and Corbett found that as stakes increased, teachers used test results to adjust instruction to a greater extent. The adjustments, unfortunately, were not viewed by teachers as substantive, but "game-like" (1991, p. 36) and aimed at raising scores, not necessarily improving student understanding.

Marso and Pigge (1992) surveyed teachers to determine the extent and effectiveness of the use of standardized test results. Both elementary and secondary teachers were asked to rate aspects of standardized test use, including use of test results

for "planning day-to-day instruction" (p. 27). The authors found that elementary teacher responses were more diverse, indicating less consistent practices related to using test results. Secondary teachers consistently reported less use of test results. One important finding in this study was lack of concerted effort to interpret test scores or discuss results on the school or department level.

Another study that examined both elementary and secondary teachers was published in 1992 by Nolen et al. A survey of Arizona teachers and administrators revealed that 38.3% of elementary teachers and 19.4% of secondary teachers use test scores to guide instruction. When administrators were surveyed, 40.1% of elementary administrators and 32.3% of secondary administrators indicated that teachers use test results for guiding instruction. The discrepancy between actual use of scores by secondary teachers and administrators perceived use was particularly alarming to the authors. They concluded that assumptions made by policymakers and administrators that teachers use test results to inform instructional decisions was not supported.

Yakimowski (1996) provided a summary of practices concerning the impact of district-selected performance assessments in four states: California, Colorado, Connecticut, and Illinois. This study did not survey teachers, but surveyed district testing personnel concerning use of standardized tests. A standard set of survey questions was used in each of the states, with another set of questions specific to each state added to the survey. The most specific questions concerning instruction asked respondents to rate the influence that assessment plays in instructional decisions. The mean Likert rating was reported for each state, ranging from 3.22 to 3.79 on a 5-point scale. While there were

statistically significant differences in ratings by state, the overall response was similar and indicated that district testing personnel perceived that instruction had been moderately influenced by using tests. However, the author proposed further research to determine actual use through teacher surveys.

Implementation of the Virginia Standards of Learning prompted McMillan et al. (1999) to investigate use of test results for instructional adjustments. Elementary and secondary teachers were included in the survey. After the first year of test implementation, the authors found that 51% of secondary teachers reported the impact as "none" or "very little." These same descriptions were selected by only 22% of elementary teachers. Comments from elementary teachers suggest that content and pace of instruction were impacted rather than the mode of instruction. Secondary teachers that reported changing instruction cited narrowing the content as the most frequent change. Conclusions of this study included lack of use by secondary teachers, and limited use by elementary teachers to make substantive changes to both content and method of instruction. The authors noted that a major limitation of this study was the small response rate of 16%, limiting generalizability of the study.

Other Important Issues Related to Teachers and Tests

Other issues related to teacher use of test scores were found in this review and provide possible explanation for the results. These issues are teachers' ability to interpret test results, teacher confidence in test validity, and test preparation methods and pressure to prepare students for tests.

Three studies (Green & Williams, 1989; Marso & Pigge, 1992; Yakimowski, 1996) found that teachers had little training or experience in interpreting test scores. This was cited as an area of concern, and a possible reason that teacher use of test scores was low. Green and Williams (1989) asked teachers to report the amount of training received in measurement and found there was a statistically significant difference in use of standardized test results based on amount of training. This issue was the topic of a review of literature by Daniel and King (1998) in which the authors found a lack of preservice or inservice training about testing and measurement.

Teacher confidence in a test's ability to validly measure student achievement was discussed in five of the seven studies. Although this issue was not a primary research focus, the studies by Nolen et al. (1992) and Yakimowski (1996) report that teachers lacked confidence in the results and may have felt that adjustments to instruction were not warranted. Wilson and Corbett (1991) reported that teachers in Maryland had more confidence in validity of test results than teachers in Pennsylvania, and Maryland teachers reported greater use of test results. It is unclear, however, whether use was due to confidence in test results or the high stakes nature of the Maryland test.

Pressure to prepare students for tests was an issue discussed in four studies. Three of these four studies pertained to high stakes tests (McMillan et al., 1999; Nolen et al., 1992; Wilson & Corbett, 1991). Teachers reported that pressure was applied by administrators, parents, and/or them due to importance of test scores. The fourth study to discuss test preparation pressure was Salmon-Cox (1981). This study discussed the increase in standardized testing for accountability and the accompanying pressure to raise

test scores. Methods used to prepare students, as cited in these studies, include review of content, review of test format, vocabulary drills, and direct teaching of test-taking skills. Nolen et al. (1991) was particularly critical of these practices and the possibility of test score pollution.

Review Conclusions

Based on the seven studies reviewed here, reported use of test scores for instructional decision making follows some general trends. First, teachers reported using results of high stakes test to a greater extent than low stakes tests. Second, elementary teachers used test results to a greater extent than secondary teachers, although use was moderate. Third, tests developed by states or selected by districts were used more extensively than national tests of achievement. It is unclear, however, whether this pattern is due to test development and selection, or level of stakes. The studies using state-developed tests also reported higher stakes and were published more recently than the studies focusing on national tests.

Based on the reviewed articles, average teacher use of tests would be rated as 1.0, using the defined metric. This indicates a moderate use of tests, with greater use reported in states with higher stakes tests. This low use is alarming in light of the assumption that such use exists and leads to higher student achievement.

Summary

The educational accountability movement is leading to increasingly higher stakes

testing by states of student achievement. Scores from state tests are used to make decisions about student achievement, qualification for graduation or promotion, effectiveness of teachers, school quality, and program effectiveness. Unfortunately, most state tests were constructed to measure the knowledge a student had acquired, not for the accountability purposes for which they are now regularly used (Barton, 1999).

In states in which test scores are used to make decisions about mastery of skills and knowledge, an important question is whether decisions made on the basis of students' test scores are appropriate and accurate. The evaluative argument approach structures validity studies to (a) identify assumptions made about test scores and score use, and (b) use methods of data collection required to address each assumption. Methods for collecting evidence for each assumption are well established. Unfortunately, with respect to tests used in state accountability programs, such investigations did not provide sufficient evidence to determine if decisions about student achievement were valid. Evidence of teacher use of test scores to make instructional decisions has not been collected by states despite the argument that scores are used to adjust instruction, leading to higher student achievement.

Utah, like the other states reviewed, provided insufficient evidence to support the argument that the UCAP measures pre-algebra knowledge and skill or the argument that teachers use UCAP scores to make instructional decisions. The evidence provided by the state relied on the test development process and did not address the underlying assumptions of the test's use. Using an evaluative argument approach, this study was designed to collect evidence addressing five of six assumptions about the UCAP. This

study represented an initial investigation of evidence to support use of the UCAP in making decisions about student mastery of pre-algebra and the extent to which teachers used scores for making instructional decisions. The stakes of decisions made based on student UCAP scores are increasing, making the investigation of validity evidence imperative.

CHAPTER III

METHODOLOGY

Purpose and Objectives

The structure of this study was based on the following arguments and assumptions:

I. UCAP test scores reflect seventh-grade students' mastery of knowledge and skill in pre-algebra.

Assumption 1: UCAP content is relevant to and representative of the Utah Core Curriculum for Pre-algebra.

Assumption 2: Answering UCAP items requires knowledge and skills of pre-algebra mathematics and is therefore considered a measure of pre-algebra.

II. Pre-algebra teachers use UCAP results to adjust instruction, leading to higher student achievement.

Assumption 3: Students' performance on the UCAP depends directly on the content and quality of instruction received.

Assumption 4: Results are provided to teachers in a timely and meaningful report.

Assumption 5: UCAP scores are properly interpreted by teachers.

Assumption 6: Teachers make appropriate and meaningful instructional adjustments based on UCAP scores.

The purpose of this study was to determine if there was evidence to support these assumptions, with the exception of Assumption 3, which was excluded due to limited time

and resources. A summary of arguments, assumptions, and methods used in this study is contained in Table 10.

Population and Sample

This study used data obtained from a suburban northern Utah school district with

Table 10

Summary of Evaluative Argument and Methods Used

Argument	Assumptions	Methods
UCAP test scores indicate 7th-grade students' level of mastery of knowledge and skill in pre-algebra.	Assumption 1. UCAP content is relevant to and representative of the Utah Core Curriculum for Pre-Algebra.	Reliability Analysis of content objective- UCAP item match
	Assumption 2. Answering UCAP items correctly requires knowledge and skills of pre-algebra mathematics and is therefore considered a measure of pre-algebra.	Convergent and discriminant correlation pattern analysis and tests of statistical significance
Pre-algebra teachers use UCAP results to adjust instruction, leading to higher student achievement.	Assumption 3. Students' performance on the UCAP depends directly on the content and quality of instruction received.	Not included in this study
	Assumption 4. Results are provided to teachers in a timely and meaningful report.	Teacher interview content analysis
	Assumption 5. UCAP scores are properly interpreted by teachers	Teacher interview content analysis
	Assumption 6. Teachers make appropriate and meaningful instructional adjustments based on UCAP scores.	Teacher interview content analysis

an enrollment from kindergarten through 12th grade of approximately 28,000 students. The district, in compliance with the USTP, administers state-developed criterion-referenced EOL tests in language arts, science, and mathematics at Grades 1 through 6; mathematics and science are tested in Grades 7 through 12. The district also annually administers the SAT-9 at Grades 3, 5, 8, and 11.

The population to which the results of this study apply are students who take the UCAP during May of their seventh-grade year, and the SAT-9 in September of their eighth-grade year.

Student Sample

Students whose data were included in this study were all seventh graders during the 1998-1999 school year. During May of 1999, students completed the UCAP as part of their pre-algebra course. In September of 1999, these same students, as eighth graders, completed the SAT-9. All students who had useable data for each of the above tests, and grades assigned for the pre-algebra course were included in the research sample (N = 1,461). Table 11 displays the student sample characteristics, and total school population characteristics for the district and state (USOE, 1999).

Teacher Sample

Pre-algebra teachers within the participating district were interviewed concerning their use of the UCAP scores. Only teachers who taught pre-algebra during the 1998-1999 school year and were teaching pre-algebra during the 1999-2000 year were included in the

Table 11

Student Characteristics for Sample and State of Utah

	Sample		District[a]		State of Utah[a]	
Characteristic	n	%	n	%	n	%
Gender						
Female	806	55.2	13,649	48.7	232,889	48.4
Male	655	44.8	14,409	51.4	148,287	51.6
Ethnicity						
American Indian	5	0.3	150	0.5	7,257	1.5
Asian/Pacific Islander	16	1.1	428	1.5	12,149	2.5
Black	11	0.7	292	1.0	3,908	0.8
Hispanic	58	4.0	1,316	4.7	34,186	7.2
White	1,371	93.8	25,872	92.2	419,561	87.9

[a] Number and percent of total school enrollment.

study (n = 12). One to three teachers from each of the district's eight junior high schools participated. Table 12 displays the characteristics of the teacher sample, including gender, years of experience, and sections of pre-algebra taught per day during the study year. The majority of teachers were female with less than 10 years of teaching experience. A description of the interview development process is included later in this chapter.

Measures

Data for this study included the student scores from the May 1999 administration of the UCAP, the September 1999 administration of the SAT-9, pre-algebra course grades, teacher rating of a subset of students as masters and nonmasters of the pre-algebra

Table 12

Teacher Sample Characteristics

Characteristic	n	%
Gender		
Female	9	75.0
Male	3	25.0
Teaching experience (years)		
1 - 5	2	16.7
6 - 10	5	41.7
11 - 15	2	16.7
16 - 20	1	8.3
21 - 25	1	8.3
26 +	1	8.3
Sections of pre-algebra taught (per day)		
1 - 3	5	41.7
4 - 5	7	58.3

content, and teacher interview responses. Three separate data sets containing test and grade information were obtained from the district after receiving permission according to district policy. These data sets were edited and combined to form the final data set used for analysis. The ratings of students as master and nonmaster were obtained during the teacher interview and added to the final data set.

UCAP

The second edition of the UCAP was developed in 1997. Its primary purpose was to (a) assess student mastery of the pre-algebra core curriculum, and (b) inform teachers of strength and weaknesses of instruction of the pre-algebra core curriculum (USOE,

1997). The test consists of 80 multiple-choice items, written to measure the core content standards. The number in parentheses indicates the items contained in each core standard subtest: number/number relationships (14), number systems/number theory (12), computation/estimation (14), patterns/functions (11), algebra (15), statistics (8), probability (5), geometry (10), and measurement (6). See Appendix C for a match of subtest items to core objectives. In addition, the items were written to require one of three mathematics skills: procedural, conceptual, or application. A limited number of items (14) were used to measure more than one core content standard. Scores from the test were reported as percent (%) correct for each standard and skill subtest. Appendix D contains a state summary report form. Individual student reports also included class, school, district, and state group averages.

Pre-algebra teachers are responsible for group-administering the UCAP to students within the pre-algebra classroom setting during May of each school year. The administration manual states that students are allowed up to two class periods to complete the test, however it is not a timed test (USOE, 1997). Student responses are recorded on a machine-readable answer sheet. These are submitted by the district to USOE for scoring. Reports of individual student performance, district performance, and state performance are returned to the district in August of each year.

Test Development

Items for the UCAP were written by a committee of 5 to 10 Utah pre-algebra teachers (IBRIC, 1999). The item content was then reviewed by a panel of mathematics

teachers, measurement specialists, administrators, and university mathematics education specialists. The review panel rated the degree to which each item matched the core standard for which it was written. After item revisions were made based on this review, tests were piloted in Utah school districts. At least three districts, and approximately 1,500 students were included in each pilot test (H. Sanderson, USOE, personal communication, May 1999). Details concerning the number of items developed, revised, and included in the final version were not included in the technical manual. Teachers involved in the pilot were asked to submit comments about item content, administration procedures, and test length. Details of the comments and changes that resulted were not included in the technical manual. Items were revised based on the teacher comments and item statistics, which included item difficulty and point-biserial correlation. The final version of the test was printed by the district for use in the state assessment program.

Reliability

The Utah Technical Manual (IBRIC, 1999) reports KR_{20} internal consistency reliability coefficients for all mathematics core tests administered. These were calculated using scores obtained from the 1999 statewide administration. The reported value for the UCAP was 0.94. Subtest reliability values were not reported; however, these have been calculated for the sample in this study and are reported in Chapter IV.

Validity

Evidence about validity of the UCAP for purposes of improving instruction have been limited to item content review and item statistics (IBRIC, 1999). This evidence was

deemed acceptable by a legislatively mandated evaluation conducted by WestEd. However, a review of literature clearly supports collection of more substantial evidence to determine if the UCAP is useful for the purpose for which it was developed.

SAT-9

The SAT-9 has been used widely as a measure of student achievement in reading, language, mathematics, science, and social science (Haladyna et al., 1998; Harcourt Brace Educational Measurement [HBEM], 1997a). The SAT-9, Advanced II, designed for eighth-grade administration consists of 338 items. The following list indicates the subtests, the number of items per subtest is indicated in parentheses: reading vocabulary (30), reading comprehension (54), math-problem solving (50), math procedures (30), prewriting (15), composing (15), editing (24), science (40), social science (40), and listening (40).

The participating district group administers the SAT-9 in September of each school year. Student responses are recorded on machine-readable answer sheets. These are submitted by the district to USOE for scoring. Reports of individual student, district, and state performance are returned to the district in January of each year. Student scores for the SAT-9 are reported to the district as raw scores, standard scaled scores, grade equivalence scores, normal curve equivalence scores, national percentile, and stanine scores. For this study students' standard scaled scores were used.

Test Development

The development of the SAT-9 involved an extensive review process, preparation

of test specifications, item development and review, national item analysis, and development of the final forms (HBEM, 1997b).

Major textbooks, state and district curricula, and trends in national standards for subject areas were used to determine the important content topics and skills to be included on the SAT-9. Once identified, these topics formed the framework of the test specifications. The blueprint for each content area outlined the topics that should be addressed, the objectives associated with each topic, and the proportion of the test that would be devoted to each (HBEM, 1997b).

Items were developed to follow the test specifications. Once developed, items were reviewed by: (a) content experts who focused on the correctness of item content, (b) editors who attended to the grammatical structure and wording of the item, (c) measurement specialists who reviewed the application of item-writing techniques, and (d) teachers who participated in the local and national tryout programs (HBEM, 1997b).

The national item tryout sample of students was selected to be representative of the national school population. These students participated by taking the SAT-9 multiple-choice tryout test, and completed some items from the SAT-8 edition for equating purposes. Item statistics were generated from the tryout results.

Based on results of steps described above, final forms of the test were developed. The technical manual for the SAT-9 lists the following criteria for item selection: (a) appropriate content fit to the test blueprint; (b) appropriate difficulty for the intended grade, and the increase or decrease in difficulty for adjacent lower or higher grades; (c) good discrimination between high scorers and low scorers (biserial correlation coefficient);

(d) appropriate clarity and interest; (e) absence of bias according to Advisory Panel and statistical procedures; and (f) good spread of students choosing each distractor (HBEM, 1997b, p. 20). Details concerning the number of items developed, revised, and included in the final version of the SAT-9 were not included in the technical manual.

Reliability

A test's reliability is the extent to which it yields consistent results. On the subtests (reading, mathematics, science, social science, and listening), each containing 40 or more items, the KR_{20} values range from 0.79 on the social science subtest to 0.91 on the reading subtest (HBEM, 1997b). The SA form of the SAT-9 contains three subtests within the language subtest. These small subtests consist of 15-24 items. The KR_{20} values for these were lower, as would be expected for subtests with fewer items. The values range from 0.57, for the 15-item prewriting subtest, to 0.65, for the 24-item editing subtest.

Validity

Validity evidence presented by the test developer was structured by "validity types." Three aspects of validity were addressed in the technical manual for the SAT-9: content, criterion-related (concurrent and predictive), and construct (HBEM, 1997b).

Content validity. Validity concerning the content of the SAT-9 was established through careful examination of the items to various content sources, such as textbooks and curriculum frameworks for the content areas. However, Harcourt Brace stated that "comparison of the content of the Stanford 9 series with the instructional objectives of a

school's curriculum will provide evidence of the validity of the Stanford for use in that school" (HBEM, 1997b, p. 43). This is appropriate advice considering use of the SAT-9 by states with varying purposes.

Criterion-related validity. To establish criterion-related validity, the SAT-9 test developers examined completion rates for students for all multiple-choice subtests at every level of the test. Additionally, a cross-sectional analysis of difficulty for students at differing points in the instructional sequence was used to show that items were more difficult for beginning students and easier for students that had received more instruction. Median biserial correlation coefficients are used to show the extent to which items and subtests separate high-scoring students from low-scoring students.

Construct validity. Correlations between the SAT-9 and another measure of achievement were used as evidence of construct validity. The Otis-Lennon School Ability Tests, Seventh Edition (also developed by Harcourt Brace) were used for this analysis. The correlation coefficient between subtests of the SAT-9 and the Otis-Lennon School Ability test range from 0.63 to 0.80.

Pre-Algebra Course Grades

Course grades are often used as a measure of acquisition of content knowledge and skill, but they are also influenced by student behaviors. Nonacademic factors that influence grading in some classrooms include attendance, tardiness, disruptions of class, class participation, respect for the teacher, or daily preparedness (Canady & Hotchkiss, 1989; Hills, 1991). Although grades are not perfect measures of mathematics achievement,

they do provide information about the student's overall performance in the pre-algebra course. Grades were used as a measure of acquisition of pre-algebra knowledge for all students in the study.

Teacher Ratings of Masters and Nonmasters of Pre-Algebra

Teacher ratings of student mathematics achievement are less likely to be contaminated by variables such as preparedness or student behavior (Hills, 1991). Ratings were used to more accurately identify a subset of students as masters and nonmasters of the content objectives. During teacher interviews, each teacher was asked to identify approximately five masters and five nonmasters in his or her class according to provided definitions. Masters were defined as students expected to have performed very well on the UCAP due to a complete and thorough knowledge of the content and skills of pre-algebra, even in the absence of one or more of the following classroom behaviors: attendance, punctuality, class participation, respect for the teacher, daily preparedness, and/or completion of homework. Nonmasters were defined as students expected to have performed very poorly on the UCAP due to a lack of knowledge of the content and skills of pre-algebra, regardless of one or more of the following classroom behaviors: attendance, punctuality, class participation, respect for the teacher, daily preparedness, and/or completion of homework. A variable containing a designation for masters ($n = 71$) and nonmasters ($n = 64$) was added to the final data set.

Teacher Interview

A semistructured teacher interview was developed to determine the degree to which teachers use UCAP scores to make instructional decisions. The interview questions were written by the researcher and based on the review of literature concerning teacher use of test scores. Questions were piloted with five mathematics teachers within the participating district who were not involved in the study, but had administered a Utah EOC tests for their teaching assignment (e.g., geometry teachers). Questions concerning adjustments to instruction were asked using open-ended questions to allow for explanation and elaboration by teachers. Other issues, such as timeliness of reports, pressure to prepare students for the test, and confidence in the test as a valid measure of mathematics were asked using closed-response Likert scales. Interviews, lasting approximately 30 minutes, were conducted by the researcher at the school of each teacher. See Appendix E for the interview protocol.

Data Preparation

Three separate data sets were obtained for this study from the participating district containing UCAP data, SAT-9 data, and pre-algebra course grade data. All three sets contained data in space-delimited ASCII format saved on computer disks. Common student variables in each data set were (a) school identification number, (b) name, (c) student identification number, (d) gender, and (e) ethnicity. These variables were used to match records from the three data sets. Student records that were not present in all three

sets were not included in the final data set. Interview data were prepared for analysis by transcription of audio-taped recordings of each interview.

UCAP Data Set

The UCAP data set contained scores for each subtest and the overall test. The individual test item responses were included in the set and recorded as letters (A, B, C, D, or E). To prepare data for analysis, letter responses were transformed to 1 or 0, representing a correct or incorrect response, respectively.

SAT-9 Data Set

The SAT-9 data set also contained student scores for each subtest. However, the scores were reported in several ways: raw score, standard score, grade equivalents, normal cure equivalents, national percentile, and stanine. The standard scaled score was used in this study, and other values were deleted from the final data file. Although individual responses were included, no item analyses were completed in this study, and these variables were removed.

One aspect of the analysis required that the reading proficiency score be calculated for each student. Harcourt Brace has defined performance standards representing a criterion-referenced interpretation of the SAT-9 subtest standard scores, and were provided in the test publisher norms book (HBEM, 1997a). Each reading proficiency level, with 1 indicating poor reading proficiency and 4 indicating excellent proficiency, represents a range of reading subtest standard scores. These ranges were used to assign

the reading proficiency level to each student record based on the reading subtest standard score contained in the data set.

Pre-Algebra Course Grade Data Set

Each student record in the data set contained letter grades assigned for each term of the course. Letter grades were transformed to numeric values using the transformation function in the SPSS software. The transformation was based on a 4-point scale (i.e., A = 4.00, A- = 3.67, B+ = 3.33,...F = 0.00). An average of quarter grades was calculated and used in the analyses.

Analysis

Analysis of Test Content: Assumption 1

Evidence collected to address content of the UCAP included item/objective match and alpha coefficient reliability analysis using the Spearman Brown prophecy formula as described in Chapter II. Inspection of item match to course objectives was completed. Judgment about the extent to which items represented course objectives was based on the opinion of the researcher. Alpha coefficient reliability estimates for UCAP subtests were calculated using the study data set, and compared after adjusting for differing number of items.

Analysis of Correlational Data: Assumption 2

Convergence of indicators as validity evidence is discussed by Messick (1993) to mean that persons who score high on the test of interest should score high on other presumed indicators of the construct being measured. Situational and method variables might influence one indicator differently from others, so it is usually better to base inferences of convergence on a combination of several indicators, preferably derived from quite different measurement methods. For this reason, UCAP scores were correlated with another multiple-choice-format test of mathematics (SAT-9) and with other measures derived by different methods, namely, pre-algebra course grades and teacher ratings of students as masters and nonmasters of pre-algebra knowledge and skills.

In addition to convergent evidence there was a need for discrimination of the construct of interest (in this study, mathematics achievement) and other constructs that should not be highly correlated. This study correlated scores from the UCAP with SAT-9 subtest scores other than mathematics: reading, language, science, social science, and listening.

Correlation Patterns

Analysis of the correlation coefficients was based on (a) examination of correlation patterns for convergent evidence, and (b) a distinction between convergent and discriminant correlation coefficients for each mathematics measure. Support for the UCAP as a measure of mathematics was determined by both types of correlation patterns. The pattern of correlation for convergent measures was not sufficient evidence that the UCAP

could be used to make decisions about student mastery of pre-algebra; lower correlations for discriminant measures were also needed.

The correlation pattern of student scores for convergent and discriminant measures was analyzed for the total sample (N = 1,461) and for subgroups. Subgroups were based on demographic characteristics (gender and ethnicity) and academic ability (reading proficiency and teacher rating of pre-algebra mastery).

Convergent measures. The UCAP was expected to have strong association with the SAT-9 Math subtest, pre-algebra course grades, and teacher rating of mastery. The pattern of the predicted correlations is shown in Table 13. The symbols represent the magnitude of the correlations, with the UCAP correlating to a greater degree with the SAT-9 Math subtest and teacher rating. The course grades were expected to have a lower correlation due to pollution by behavioral factors other than math achievement, as described in the review of literature. This pattern was expected to be present for the sample and all subgroups.

Distinction of convergent and discriminant measures. The UCAP correlation to convergent measures was predicted to exceed the UCAP correlation to discriminant measures. Because constructs measured by the discriminant SAT-9 subtests may overlap mathematics, the UCAP correlation with discriminant measures was expected to be positive, but not to exceed the correlation with the convergent measures. This pattern was expected for the sample and all subgroups. Table 13 displays the predicted correlation coefficient patterns contrasting the convergent and discriminant measures. The patterns of correlation coefficients, not actual values, are indicated in the table.

Table 13

Expected Correlation Patterns for Convergent and Discriminant Validity Evidence

Measure	1	2	3	4
Convergent				
1. UCAP total	--			
2. SAT-9 math	***	–	--	--
3. Course grade	**	**	***	
4. Master/nonmaster rating	***	***		
Discriminant				
5. SAT-9 reading	*	*	*	*
6. SAT-9 language	*	*	*	*
7. SAT-9 science	*	*	*	*
8. SAT-9 social science	*	*	*	*
9. SAT-9 listening	*	*	*	*

Note. UCAP = Utah Core Assessment Pre-Algebra; SAT-9 = Stanford Achievement Test, 9th Edition.

Statistical Significance of Correlational Data

Statistical significance of correlation coefficients is influenced by sample size. Due to the large sample used in this study, analysis of correlation coefficient significance was interpreted with this in mind. Significance testing was also used to determine if the expected pattern of correlation coefficients was supported. Meng, Rosenthal, and Rubin (1992) described tests for comparing correlation coefficients. The equations presented by Meng et al. were used to determine if (a) the pattern of convergent correlation coefficients supported the expected pattern (e.g., that teacher rating was more highly correlated to the UCAP than course grade), and (b) the differences between convergent and discriminant correlation coefficients were statistically significant.

The following equation was used to make pairwise comparisons of convergent correlation coefficients to test if predicted patterns were supported by the data. It yields a $\underline{Z}$ test for the significance of the difference between two sample correlation coefficients r_{yx1} and r_{yx2} where variables x_1 and x_2 are other measures of mathematics and variable y is the UCAP score.

$$Z = (z_{r1} - z_{r2})\sqrt{\frac{N-3}{2(1-r_x)h}} \quad , \tag{3.1}$$

where $\underline{N}$ is the number of subjects, z_{ri} is the Fisher z-transformed $r_i = r_{yxi}$, and r_x is the correlation between the two predictor variables x_1 and x_2 (i.e., correlation between the course grade and SAT-9 math score). The equations for h is:

$$h = \frac{1 - f\overline{r^2}}{1 - \overline{r^2}} = 1 + \frac{\overline{r^2}}{1 - \overline{r^2}}(1 - f) \quad , \tag{3.2}$$

where

$$f = \frac{1 - r_x}{2(1 - \overline{r^2})} \quad . \tag{3.3}$$

In these equations, $\overline{r^2}$ is the mean of the squared correlation coefficients, $(r_1^2 + r_2^2)/2$.

To analyze the difference between the convergent and discriminant correlation coefficients, two tests were completed. First, a test of the homogeneity of UCAP

correlations with both convergent and discriminant measures was applied. The null hypothesis for this test stated that all correlation coefficients were equal. The equation is:

$$x^2(k-1) = \frac{(N-3)\sum_i (z_{r_i} - \overline{z_r})^2}{(1-r_x)h} \qquad (3.4)$$

where $\overline{z_r}$ is the mean of all z-transformed correlations. The resulting chi-square statistic is distributed on k-1 degrees of freedom where k is the number of correlations being tested for homogeneity. In this study, the number of correlations is eight (SAT-9 math, course grade, teacher rating, and SAT-9 subtest scores for reading, language, science, social science, and listening). The value of r^2 is the average of all k values of r_i^2, and r_x is the median intercorrelation among the variables being tested.

Finally, if the null hypothesis was rejected for the above test of homogeneity, a test of contrast was used. This allowed the convergent correlation coefficients to be contrasted with the discriminant correlations. Contrast weights represented as lambdas (λ) were assigned to each coefficient. Convergent coefficients were assigned positive lambdas, while discriminant coefficients were assigned negative lambdas such that the sum of all lambdas equaled zero. The null hypothesis for this test stated that the set of convergent correlation coefficients was equal to the set of discriminant correlation coefficients. The equation uses the result of the previous chi-square test (Equation 3.4) and the correlation coefficient between the lambdas of the contrast weights and their corresponding z_{ri}s (i.e., $r_{\lambda z}$).

$$Z = r_{\lambda z_r} \sqrt{x^2 (k - 1)} \tag{3.5}$$

Rejecting the null hypothesis for this test indicated that there was a statistically significant difference between the sets of convergent and discriminant correlation coefficients, supporting the assumption that the UCAP was a measure of pre-algebra.

Analysis of Teacher Interview Data: Assumptions 4-6

Content analysis of teacher responses to open-ended interview questions were used to ascertain the degree to which teachers made adjustments to the instruction of the pre-algebra curriculum, and the nature of those changes. Tape recordings of each interview were transcribed and segmented into units. A unit was an interview question and corresponding responses. Common responses or themes emerged from the responses of open-ended questions. These themes were then used to report and discuss the interview results.

Closed response questions that used a Likert scale were quantified to allow for general analysis. The percentage of teachers selecting each Likert point was reported. Additional comments made by the teachers and responses to follow-up questions for the closed-response questions were also analyzed for content in the same manner as described for the open-response questions.

CHAPTER IV

RESULTS

Validity is the degree to which a test accomplishes what it was designed to accomplish. Tests are not classified as "valid" or "invalid," but rather as possessing degrees of validity for different purposes. This study sought validity evidence for the assumptions underlying the purposes of the UCAP to determine if the test could be used to make valid decisions about students' mathematics achievement. This chapter presents findings for each assumption for which evidence was collected.

Evidence for Assumption 1

Evidence collected to support Assumption 1 included the review of UCAP items to course objective, reliability estimates, and subtest intercorrelations.

Match of UCAP Items to Pre-Algebra Objectives

IBRIC (1999) cited the test development process as evidence that UCAP items were relevant and representative of pre-algebra course objectives. However, the lack of detail in the description of the process, as described in the technical manual, led to the need to analyze the item-objective correspondence. Table 14 displays a summary of this match. Appendix B contains the match of pre-algebra course objectives and item number. Neither the number of objectives in each subtest nor the percentage of objectives measured in each subtest reflect a "weight" of importance or content emphasis. The

Table 14

UCAP Subtest, Objective, and Item Correspondence

UCAP subtest	# of objectives	# of items	# of objectives measured	% of objectives measured
Number/number relationships	5	14	5	100
Number systems/theory	5	12	3	60
Computation/estimation	6	14	4	67
Patterns and functions	4	11	3	75
Algebra	6	15	5	83
Statistics	5	8	2	40
Probability	5	5	3	60
Geometry	6	10	3	50
Measurement	7	6	4	57
Total test	49	95[a]	32	65

[a]Some items are assigned to more than one subtest.

number of items assigned to each subtest appeared to provide a weight, although the overall percentage of objectives measured is moderate, 65%. Upon further inspection of the item to objective match, it was found that 49 of the 80 items are assigned to 10 objectives. The UCAP did not contain items that were representative of the entire pre-algebra core curriculum.

Reliability

Measures of internal consistency included both reliability coefficients and

intercorrelation of subtest scores with the total test score. The UCAP contains nine content subtests and three skill subtests. The reliability coefficients and intercorrelations of these subtests are presented and discussed in this section. The coefficient of internal consistency provides an estimate of how consistently examinees perform across items within a test during a single test administration. USOE conducted reliability studies of the internal consistency of the UCAP using the 1999 statewide administration ($\underline{N}$ = 29,944). The KR_{20} internal consistency coefficient of the UCAP was reported as 0.94 (IBRIC, 1999).

As discussed in Chapter II, reliability coefficients are sensitive to the number of items contained on the test (Crocker & Algina, 1986), and therefore the coefficients of UCAP subtests containing different numbers of items cannot be directly compared. To overcome this problem, the Spearman Brown prophecy formula was employed to obtain the adjusted estimate of the reliability coefficient of the UCAP subtests as if each had the same number of items. Table 15 displays the original and adjusted alpha coefficients for the UCAP subtests as calculated using test results of this study sample ($\underline{N}$ = 1,461). After adjustment for number of items, the subtest reliability coefficients ranged from 0.82 to 0.93. The values were compared to the SAT-9 mathematics subtest reliability coefficients to determine if they were consistent with another measure of mathematics achievement. The Spearman Brown adjustment was made to the SAT-9 reliability coefficients in the same manner described for the UCAP, using 32 items in the calculation of the K, the ratio of the original and new number of test items. Table 16 displays the original and adjusted values of the SAT-9 mathematics subtest reliability coefficients. These adjusted values

Table 15

Original and Adjusted Reliability Coefficients for the UCAP Content and Skills Subtests

UCAP subtest	# of items	Original coefficient alpha	Adjusted coefficient alpha
Content subtests			
Number /number relationships	14	0.72	0.85
Number systems /number theory	12	0.67	0.84
Computation and estimation	14	0.69	0.84
Patterns and functions	11	0.66	0.85
Algebra	15	0.74	0.86
Statistics	8	0.69	0.90
Probability	5	0.68	0.93
Geometry	10	0.62	0.84
Measurement	6	0.82	0.86
Skill Subtests			
Procedural	25	0.82	0.86
Conceptual	32	0.82	0.82
Application	19	0.79	0.86
Total test	80	0.93	0.84

Note. Adjusted coefficient alpha was calculated using the Spearman Brown Prophecy formula. UCAP = Utah Core Assessment Pre-Algebra.

indicated that the UCAP subtest reliability coefficients were very similar to the SAT-9 mathematics subtest reliability coefficients.

Intercorrelation of UCAP Subtests

Another means of determining the internal consistency of the UCAP was to correlate the subtest scores with the total test score. These coefficients provide evidence of the UCAP's construct validity, although it is essential that this internal validation be

Table 16

Original and Adjusted Reliability Coefficients for the SAT-9 Mathematics Subtests

Measure	# of items	Original alpha coefficient	Adjusted alpha coefficient
SAT-9 Mathematics			
Procedures	30	0.82	0.83
Problem Solving	50	0.86	0.80
Mathematics Total	80	0.91	0.80

Note. Adjusted coefficient was calculated using the Spearman Brown Prophecy formula. SAT-9 = Stanford Achievement Test, 9th Edition.

complemented by validity studies that employ external criteria (Anastasi, 1982). Table 17 presents internal consistency correlation coefficients for the UCAP total score and content subtests for the total sample (N = 1,461). Presented in Table 18 are the UCAP skill subtest correlation coefficients for the total sample. The subtest intercorrelations are moderate (.41 to .76). The weakest relationships were found with the statistics, probability, and measurement subtests; however, the correlations are still considered moderate with a range of .41 to .61.

The highest intercorrelations are found among the first five subtests, which contain 82.5% of the UCAP items. These values range between .58 to .76. The intercorrelations of the UCAP content subtests and total score provided evidence that the subtests measure similar constructs. The correlation of subtests to total test ranged from .67 to .86, indicating strong relationships between subtests and the total test.

Table 17

Intercorrelation of UCAP Content Subtests and Total Score

UCAP subtest	1	2	3	4	5	6	7	8	9
1. Number and number relationships	--								
2. Number systems/number theory	.68	--							
3. Computation and estimation	.71	.67	--						
4. Patterns and functions	.66	.58	.57	--					
5. Algebra	.73	.76	.69	.66	--				
6. Statistics	.55	.49	.53	.61	.57	--			
7. Probability	.43	.41	.42	.50	.47	.61	--		
8. Geometry	.49	.44	.51	.54	.52	.62	.56	--	
9. Measurement	.54	.50	.60	.52	.54	.49	.41	.51	--
10. Total test	.84	.79	.82	.81	.86	.78	.67	.73	.70

Note. All coefficients $p < .01$. UCAP = Utah Core Assessment Pre-algebra. See Table 15 for number of items per subtest.

Table 18

Intercorrelations of UCAP Skill Subtests and Total Score

Measure	1	2	3	4
UCAP				
1. Total score	--			
2. Procedural	.91	--		
3. Conceptual	.94	.78	--	
4. Application	.90	.73	.80	--

Note. All coefficients $p < .01$ UCAP = Utah Core Assessment Pre-Algebra. See Table 15 for number of items per subtest.

Evidence for Assumption 2

To determine if UCAP scores can be used to make valid decisions about student knowledge and skill in pre-algebra, convergent and discriminant correlations were analyzed. The correlations used for convergent and discriminant evidence were examined for the total sample (N = 1,461) and for subgroups. Subgroups were based on demographic characteristics, gender and ethnicity, and academic ability, reading proficiency and teacher rating of pre-algebra mastery. Table 19 displays a crosstabulation to further define these subgroups.

Descriptive Statistics

Convergent Measures

Table 20 displays means and standard deviations of three convergent measures: UCAP total score, SAT-9 math total score and pre-algebra course grade.

Initial review of mean scores for convergent measures revealed that poor readers (Level 1) had the lowest mean UCAP score (40.60 %), while students rated as masters of pre-algebra had the highest mean UCAP score (78.96 %). The extreme high and low mean standard scores for SAT-9 math belonged to the Level 4 readers (725.68) and Level 1 readers (651.86), respectively. Not surprisingly, pre-algebra course grades were most distinct for students rated as masters and nonmasters of pre-algebra.

Means for demographic subgroups were compared for statistically significant differences using one-way analysis of variance (ANOVA). Gender subgroups had

Table 19

Crosstabulation of Subgroups with Reading Proficiency and Mastery of Pre-algebra Knowledge and Skills

	SAT-9 reading proficiency							
	Level 1		Level 2		Level 3		Level 4	
Groups	n	%	n	%	n	%	n	%
Males	59	9.0	304	46.4	271	41.4	21	3.2
Females	54	6.7	342	42.4	356	44.2	54	6.7
Minority	11	12.2	44	48.9	30	33.3	5	5.6
Majority	102	7.4	602	43.9	597	43.5	70	5.1
Master	1	1.4	10	15.6	49	68.0	11	15.3
Nonmaster	13	20.3	36	56.2	15	23.4	0	0.0

	Teacher rating			
	Nonmaster		Master	
	n	%	n	%
Males	32	4.9	21	3.2
Females	32	4.0	50	6.2
Minority	2	2.2	1	1.1
Majority	62	4.7	70	5.3

Note. Percent are calculated based on subgroup size. Reading proficiency based on Stanford Achievement Test, 9th Edition Reading Subtest.

statistically significant differences for course grade. Statistically significant differences were also found for the UCAP and SAT-9 math subtest in the ethnicity subgroup. The two subgroups based on academic achievement, reading proficiency, and teacher rating, had statistically signifiicant differences between scores on each mathematics test and the course grade. For all convergent measures, the Level 4 readers and students rated as masters of pre-algebra knowledge had higher mean scores than their subgroup

Table 20

Means and Standard Deviations of UCAP Total Score, SAT-9 Math Score, and Pre-Algebra Course Grade for the Sample and Subgroups

		UCAP total score		SAT-9 math total		Pre-algebra course grade	
Group	N	M	SD	M	SD	M	SD
Sample	1461	58.12	18.04	685.27	32.56	2.64	1.18
Gender							
Males	655	57.92	18.64	686.64	34.58	2.45	1.19
Females	806	58.27	17.55	684.16	30.79	2.79	1.14
Ethnicity							
Minority	90	51.44	18.56	676.66	32.34	2.42	1.26
Majority	1371	58.55	17.93	685.84	32.50	2.65	1.17
Reading proficiency							
Level 4	75	77.16	14.33	725.68	29.88	3.42	.93
Level 1	113	40.60	12.73	651.86	17.22	1.80	1.22
Teacher rating							
Masters	71	78.96	11.21	721.32	33.61	3.77	.40
Nonmasters	64	45.00	16.90	661.69	24.88	1.21	1.00

Note. UCAP means are percent correct; SAT-9 means are standard scores; grade based on four point scale. UCAP= Utah Core Assessment Pre-Algebra; SAT-9 = Stanford Achievement Test, 9th Edition.

counterparts. Tables 21, 22, and 23 display the results of the ANOVA for the UCAP, SAT-9, and course grade means, respectively

Standardized mean differences were calculated for each subgroup. The standardized mean difference is helpful in comparing the differences in performance of measures with different scales (Gall et al., 1996). The formula numerator was the

Table 21

One-Way ANOVA of UCAP Mean Score for Subgroups

Group	Source	df	SS	MS	F
Gender	Between groups	1	43.08	43.08	.132
	Within groups	1459	475250.37	325.74	
Ethnicity	Between groups	1	4268.42	4268.42	13.22*
	Within groups	1459	471025.03	322.84	
Reading proficiency	Between groups	1	60249.32	60249.32	336.05*
	Within groups	186	33347.16	179.29	
Teacher rating	Between groups	1	37164.75	37164.75	170.38*
	Within groups	135	29446.72	218.12	

*$p < .05$.

Table 22

One-Way ANOVA of SAT-9 Mean Score for Subgroups

Group	Source	df	SS	MS	F
Gender	Between groups	1	2212.41	2212.41	2.089
	Within groups	1459	1545445.63	1059.25	
Ethnicity	Between groups	1	7122.34	7122.34	6.75*
	Within groups	1450	1540535.70	1055.89	
Reading proficiency	Between groups	1	245668.05	245668.05	460.41*
	Within groups	186	99246.06	533.58	
Teacher rating	Between groups	1	113985.02	113985.02	122.01*
	Within groups	135	126119.391	934.22	
	Within Groups	135	126119.391	934.22	

*$p < .05$.

Table 23

One-Way ANOVA of Course Grade Mean for Subgroups

Group	Source	df	SS	MS	Group
Gender	Between groups	1	41.54	41.54	30.64*
	Within groups	1459	1978.53	1.36	
Ethnicity	Between groups	1	4.75	4.75	3.44
	Within groups	1459	2015.32	1.38	
Reading proficiency	Between groups	1	118.14	118.14	95.63*
	Within groups	186	229.78	1.24	
Teacher rating	Between groups	1	214.91	214.91	347.63*
	Within groups	135	83.46	.618	

*$p < .05$.

difference between mean scores of the two groups. The denominator is the sample standard deviation (see Table 24).

Standardized mean differences for the UCAP ranged from 0.02 for the gender subgroups to 2.03 for the reading proficiency subgroups. For the gender subgroup the standardized mean difference was very small and supported the statistically nonsignificant difference in mean scores. The ethnicity subgroup standardized mean difference was 0.40. As expected for groups intended to discriminate performance on these measures, the standardized mean difference for the academic achievement groups was extremely large, 1.88 to 2.03.

Standardized mean differences for the SAT-9 are similar to those of the UCAP, with the gender subgroup having a small value (0.08), and the ethnicity subgroup a more moderate value, 0.28. As expected, and consistent with the results for the UCAP, the

Table 24

Standardized Mean Differences of Convergent Measures for Subgroups

Groups	UCAP	SAT-9 math subtest	Grades
Gender	0.02	0.08	0.29
Ethnicity	0.40	0.28	0.20
Reading proficiency	2.03	2.27	1.37
Teacher rating	1.88	1.83	2.17

standardized mean difference for subgroups based on academic achievement was extremely large, 1.83 and 2.27.

Differences in standardized means for subgroups based on demographics are more similar for course grades than for the other two measures. The gender subgroup, which had a statistically significant mean difference, also had a more moderate standardized mean difference of 0.29. In fact, this value was greater than that of the ethnicity subgroup. For the reading proficiency subgroup the standardized mean difference was slightly smaller for course grade than for the other two measures, but still very large at 1.37. The largest standardized mean difference for course grade was found for the teacher rating subgroup, 2.17.

When analyzed by subgroup, the standardized mean differences reveal that differences in teacher grades were greater than differences of math test means for both the gender and teacher-rated subgroups. For both of these groups the standardized mean differences were similar for the UCAP and SAT-9 math subtest. Students in the

contrasting reading proficiency groups had similar standardized mean differences for the UCAP and SAT-9 math subtest, with both being greater than the standardized difference for grades. Finally, the teacher-rating subgroup had lower standardized mean differences for the two math tests than for grades.

Discriminant Measures

Table 25 presents the mean scores of the five SAT-9 subtests serving as discriminant measures: reading, language, science, social science, and listening for the sample and subgroups.

Again, one-way ANOVA was used to analyze the mean scores for statistically significant differences. Table 26 displays the results of this analysis.

Statistically significant differences were found for most tests. To further analyze these differences, standardized mean values were calculated. These values are contained in Table 27. The standardized differences for the gender subgroup are very small for the social studies and listening (0.05), and moderate for the other three tests in which there was a statistically significant difference in mean scores (0.18 to 0.32). The subgroup based on ethnicity had moderate standardized mean differences for all tests, ranging from 0.16 to 0.38. As expected, the subgroups based on academic achievement had very large standardized differences for all tests, ranging from 1.10 to 3.89.

Table 27

Standardized Mean Differences of Discriminant Measures for Subgroups

Groups	Reading	Language	Science	Social Science	Listening
Gender	0.18	0.32	0.33	0.05	0.05
Ethnicity	0.23	0.20	0.30	0.16	0.38
Reading proficiency	3.89	2.88	2.32	2.38	2.42
Teacher rating	1.35	1.56	1.22	1.10	1.10

Correlation of Convergent and Discriminant Measures

Evidence to support the use of the UCAP as a measure of student achievement was collected using convergent and discriminant measures. The correlations for these measures were analyzed in two ways: pattern of convergent measures, and distinction of convergent from discriminant measures. The results of these correlations are presented and discussed in this section.

Analysis of Convergent Measures

Patterns. Table 28 displays the results of convergent measures correlations. Convergent correlation coefficients are listed for the sample and each subgroup. Shaded cells indicate instances in which teacher rating was a constant variable and correlations could not be computed; all students who had a reading proficiency level of 4 and were rated by teachers, were rated as masters. Displayed in Appendix F, Tables F1-F5 are the correlation tables with coefficients of all measures for the sample, demographic

Table 28

Summary of Convergent Correlations for the Sample and Subgroup

		UCAP convergent correlation coefficients		
Groups	N	SAT-9 math subtest	Course grade	Teacher rating
Sample	1,461	.73	.61	.75
Females	806	.74	.62	.80
Males	655	.71	.62	.80
Majority	1,371	.73	.61	.75
Minority	90	.66	.67	1.00[a]
Reading level 1	113	.53	.42	.38
Reading level 4	75	.65	.66	
Nonmaster	64	.52	.70	
Master	71	.71	.55	

Note. Constant teacher rating indicated by shaded cells. UCAP = Utah Core Assessment Pre-Algebra; SAT-9 = Stanford Achievement Test, 9th Edition.
[a]n = 3.

subgroups of gender and ethnicity, and achievement subgroups of reading proficiency and mastery of pre-algebra.

There were two exceptions to the expected pattern. The first exception was found for students with a level 1 reading proficiency. For these students the UCAP to course grade correlation (.42) was higher than the UCAP to teacher rating correlation (.38). The second exception to the hypothesized pattern involved the correlation of UCAP with course grades that exceeded the correlation of UCAP with SAT-9 math subtest. This was the case for the ethnic minority students (.66 vs. .67), level 4 reading students (.65 vs. .66), and students rated as nonmasters (.52 vs. .70). The coefficient difference of .01 for the ethnic minority and level 4 reading students was considered very small, with little

substantive meaning. For students rated as nonmasters, however, the difference in these correlation coefficients was much greater.

Inferential tests. A test was completed to determine if there were statistically significant differences between convergent correlation coefficients, as indicated in the expected patterns. This test made pairwise comparisons using equations developed by Meng et al. (1992) and produced a Z-value, as described in Chapter III, Equation 3.1. The null hypothesis for this test stated that the correlation coefficients were equal: A rejection of the null hypothesis supported the expected pattern of coefficients. Table 29 contains the Z-values of these comparisons.

The test for differences in correlation coefficients indicates that correlation coefficients for minority students, and subgroups based on academic achievement were not statistically significant and did not support the expected pattern. The statistically significant differences between the course grade and rating supported the expected pattern and the use of ratings as a measure of student achievement, for all groups except males.

Distinction of Convergent and Discriminant Correlations

Although the test of statistically significant differences between convergent measures was used to understand the relationships between those measures, a more important piece of evidence to support the assumption that the UCAP measures pre-algebra knowledge and skill was the distinction of convergent and discriminant measures.

The expected pattern of distinctions between convergent and discriminant correlations of UCAP scores was supported for all groups except students rated as

Table 29

Z-Values for Pairwise Comparisons of Convergent Measures

Groups	UCAP convergent correlation		
	SAT-9 math subtest versus course grade	Course grade versus teacher rating	SAT-9 math subtest versus teacher rating
Sample	4.94*	3.54*	0.88
Females	3.84*	5.47*	2.23
Males	2.57*	1.36	0.69
Majority	4.84*	4.76*	0.76
Majority	0.07	13.28*[a]	13.29*[a]
Reading level 1	0.90	2.87*	1.26
Reading level 4	0.07	5.00*	NA
Nonmaster	1.34	4.99*	NA
Master	1.47	3.75*	NA

[a] Value calculated using correlation coefficient of 1.00 for $n = 3$ students.
*$p < .05$. NA = not available, teacher rating was a constant value.

masters. These students' UCAP correlation to SAT-9 subtests in reading, language, and science exceeded the UCAP correlation to the pre-algebra course grade, but not the SAT-9 math subtest.

To test if the differences between the convergent and discriminant correlation coefficients were statistically significant, two tests were performed. First, a test of homogeneity was used, yielding a chi-square value. The null hypothesis for this test states that all correlation coefficients are equal. For the sample and all subgroups, the null hypothesis was rejected. This allowed a test of contrast between the convergent and discriminant correlations to be completed, using Formula 3.5. The null hypothesis for this test states that the set of convergent correlations were equal to the discriminant

correlations as a set. The resulting Z-values led to a rejection of the null hypothesis for the sample and all groups, indicating that the set of convergent measures was statistically significantly higher than the set of discriminant measures. A summary of the contrasting convergent and discriminant correlation results are presented for all groups in Table 30. The range of correlations are shown in the summary table, with complete tables for each group displayed in Appendix F, Tables F1-F5.

Evidence for Assumptions 4 Through 6

Pre-algebra teachers in the participating district were interviewed to determine if there was evidence to support Assumptions 4-6 concerning teacher receipt of a meaningful test report, interpretation of scores, and adjustment of instruction. This section presents responses to interview questions concerning receipt, interpretation, and use of UCAP results for making instructional adjustments.

Receipt of UCAP

Before UCAP results can be used to adjust instruction, they must be received and interpreted by teachers. Several questions were asked to determine if UCAP results were received by teachers.

The types of UCAP reports sent to each school were student reports, teacher reports, and school reports. Table 31 displays the results of Questions 1 and 11 pertaining to the receipt of reports, and the timeliness of the return of results. Of the 12 teachers included in this study 50.0% reported receiving student reports, 75.0% received teacher

Table 30

Summary of Convergent Versus Discrimant Correlations for Sample and Subgroups

	Contrasting correlations			
Groups	UCAP versus convergent measures	UCAP versus discriminant measures	χ^2 values	Z-values
Sample	.61 - .75	.48 - .59	483.48*	18.60*
Females	.62 - .80	.48 - .59	268.44*	13.86*
Males	.62 - .71	.45 - .60	220.22*	12.55*
Majority	.61 - .75	.47 - .60	453.65*	18.01*
Minority	.66 - 1.00	.44 - .58	118.96*	9.22*
Reading level 1	.38 - .53	.04 - .31	324.75*	15.24*
Reading level 4	.65 - .66	.21 - .42	109.16*	8.84*
Nonmaster	.52 - .70	.33 - .51	48.82*	5.91*
Master	.55 - .71	.29 - .62	31.42*	4.74*

Note. UCAP = Utah Core Assessment Pre-algebra; SAT-9 = Stanford Achievement Test, 9th Edition.
*$p < .05$.

results, and 91.6% of teachers received school reports. The majority of teachers in this sample (75.0%) reported that they received school-level reports of UCAP results. However, more than half of the teachers receiving the results reported that the reports were not given to them directly; instead they are "invited" to view them, or sought them out independently. One teacher commented, "We were invited [by the principal] to look at them. We didn't actually get them, just looked at them." Another complained, "I went and found them! They were in the office, I had to go down and find them so we [referring to the department] could see how we did." Most teachers, 58.3%, felt that UCAP results were returned in a timely manner to facilitate instructional adjustments, but two strongly disagreed with this statement. The primary concern expressed was receiving the results in

Table 31

Teacher Responses to Question Concerning Receipt of UCAP Results

Question	Response	n	%
1. Did you receive the score reports for the May 1999 administration of the UCAP?	Yes	11	91.7
	No	1	8.3
If yes, were the results reported by student?	Yes	6	50.0
	No	5	41.7
	Not applicable	1	8.3
If yes, were the results reported by teacher?	Yes	9	75.0
	No	2	16.7
	Not applicable	1	8.3
If yes, were the results reported by school?	Yes	11	91.7
	No	0	0.0
	Not applicable	1	8.3
11. The results of the UCAP are returned in a timely manner to allow for adjustments to the instruction of the pre-algebra curriculum to be made.	Strongly agree	2	16.7
	Agree	5	41.7
	Disagree	3	25.0
	Strongly disagree	2	16.7
	Not applicable/Don't know	0	0.0

Note. Percent is based on number responding to question. UCAP = Utah Core Assessment Pre-Algebra.

the days just prior to school starting, after most instructional planning had occurred. One teacher said, "They get back late! However, I don't want to test any earlier in the year. I guess I can't have it both ways."

In order for instructional adjustments to be made, a careful examination of test results is required. When asked about careful examination of the results, however, only 1 of the 12 teachers responded that student results had been carefully examined. She commented that the purpose of examining the student reports was to confirm the placement of

students for the subsequent math course. Half the sample teachers reported a careful examination of their own class results while 75.0% carefully examined the school-level report. Less than half the teachers reported that the math department as a whole examined the school results, and discussed areas of concern. Those that did report discussing results as a department cited content of the pre-algebra course, content of test items, and logistics of administering the test as areas of concern. Table 32 contains the results of Question 2 and 3.

Table 32

Teacher Responses to Question Concerning Examination of UCAP Results

Question	Response	n	%
2. Did you carefully examine the results for each student?	Yes	1	8.3
	No	5	41.7
	Not applicable	6	50.0
Did you carefully examine the results for your own classes?	Yes	6	50.0
	No	3	25.0
	Not applicable	3	25.0
Did you carefully examine the results for the school?	Yes	9	75.0
	No	2	16.7
	Not applicable	1	8.3
3. Did your department, as a group, carefully examine the results for the school?	Yes	5	41.7
	No	5	41.7
	Not applicable	2	16.7
If yes, did the department discuss areas of concern?	Yes	5	41.7
	No	0	0.0
	Not applicable	7	58.3
If yes, what were the areas of concern?	Logistics	2	16.7
	Content	2	16.7
	Item types	1	8.3

Note. Percent is based on number responding to question. UCAP = Utah Core Assessment Pre-Algebra.

Interpretation of UCAP Scores

Teachers were asked five questions concerning the interpretation of UCAP scores. Teachers expressed confidence in their ability to interpret the results, and felt that the report facilitated necessary adjustments to the instruction of the pre-algebra content.

According to the USOE, the top portion of the report is intended to provide teachers with a percent correct score for each of the nine content subtests and the three skill subtests (USOE, 1997). A student report displays scores for student, school, district, and state. The lower section of the form provides information about student performance on specific sets of items within the subtests that are reported as raw scores. The purpose of this section is to provide teachers with enough detail to adjust instruction of content within the subtests. Appendix D contains a district level UCAP report. Teachers were asked to identify areas on a state summary report that provided information that was most helpful for determining (a) how well students had performed (Question 6), and (b)what adjustments should be made to the instruction of pre-algebra (Question 8). Table 33 displays the responses of teachers to these questions. For the purpose of determining how well students performed, 41.6% of teachers reported using the top portion as designed by USOE, while 33.3% reported using the bottom portion and 16.7% used both. This suggests that teachers are not interpreting the scores as USOE intended; however, information at the bottom of the report is more detailed than at the top, and scores are reported as raw scores. The teachers are not likely to have any misunderstanding when using both portions in determining student performance.

Table 33

Teacher Response to Questions Concerning the Interpretation of UCAP Results

Question	Response	n	%
6. Using this sample report form as an example, what information on the report is helpful to you in determining how well your students performed on the UCAP?	Summary at top	5	41.7
	Detail at bottom	4	33.3
	Both the top and bottom	2	16.7
	Don't know /Never used	1	8.3
7. Which information, the actual percent correct or the comparison to district and/or state performance, is most important to you in determining how well your students performed on the UCAP?	Actual percent correct	1	8.3
	Comparison to others	10	83.3
	Don't know	1	8.3
8. Using this sample report form as an example, what information is helpful to you in determining what adjustments should be made to your pre-algebra instruction?	Summary at top	0	0.0
	Detail at bottom	9	75.0
	Both the top and bottom	2	16.7
	Don't know /Never used	1	8.3
12. I have confidence in my ability to interpret the results of the UCAP.	Strongly agree	5	41.7
	Agree	5	41.7
	Disagree	0	0.0
	Strongly disagree	0	0.0
	Not applicable /Don't know	2	16.7
13. The information provided on the test report facilitates adjustment to my pre-algebra.	Strongly agree	3	25.0
	Agree	6	50.0
	Disagree	2	16.7
	Strongly disagree	0	0.0
	Not applicable /Don't know	1	8.3

Teachers' responses were more consistent when asked about the portion of the report used for adjusting instruction. Seventy-five percent of teachers reported using the bottom portion, as intended by the state, while 16.7% used both the top and bottom

portions. Again, the information contained in each portion of the report differs in level of detail. It is unlikely that misunderstanding could result from using both portions.

Finally, teachers were asked which was more important to them in determining how well their students performed, the percent correct or the comparison to others' performance. Nearly all teachers, 83.3%, responded that comparison to others was most important. One teacher indicated that this had become more important to her since her principal began emphasizing the school's performance, rather than the individual teacher results. She added, "I know the administration worries about the comparison because they were saying things like, 'How can we be at the bottom of the district?' They just want to look good." The one teacher that reported using the percent correct as the most important measure commented, "Comparing would make it norm referenced. I want to know how [students] did on the actual test, not just compared to other kids. If everyone is bombing the test, that doesn't matter. I worry about my students, not everyone else."

Teachers seemed capable of interpreting test scores using the UCAP reports, but they placed more importance on the comparison of student performance to others than to the actual score. The teachers were interpreting the UCAP in a norm-referenced manner, rather than criterion-referenced. This finding is alarming in light of the purpose of the UCAP to help teachers determine if students have mastered pre-algebra skills, not as a tool for comparison.

Use of UCAP to Inform Instruction

Teachers were asked seven questions concerning the use of the UCAP for making

instructional decisions and steps taken to prepare students for the test. Specific examples were elicited from teachers about adjustments made. These responses were categorized for analysis.

Instructional Adjustments

Open-ended questions were asked to determine what adjustments, if any, teachers made to their instruction based on UCAP results (see Table 34).

Question 4 asked about adjustments made during the 1999-2000 school year, based on the 1999 administration of the UCAP. Within the sample, 66.7% of teachers responded that adjustments had been made. The adjustments were grouped in three categories, and the percentage of teachers making each type of adjustment is indicated in parentheses: content (66.7 %), method of teaching (50.0%), and increased use of manipulatives or technology (25.0%). Content adjustments included adding or deleting specific content based on its presence/absence on the UCAP, particularly increasing coverage of probability and statistics, graphing, and real world applications. Teaching methods pertained to the implementation of instruction, including sequence of content presentation, emphasis on review of previous content, and alterations in the presentation of new content. During the interviews teachers explained new methods for presenting word problems, increased time for review, and the addition of review problems in the daily lesson. These adjustments were grounded in the UCAP results; however, only 41.7% of teachers gave an example of changing content in terms of a specific subtest. Review of material was cited most often, but specific details of the review methods were not given.

Table 34

Teacher Response to Question Concerning Use of UCAP Scores for Instructional Adjustment

Question	Response	n	%
4. Based on the results of the summary reports or discussions, have you made adjustments to your instruction of the pre-algebra curriculum?	Yes	8	66.7
	No	4	33.3
If yes, what were the adjustments	Content (add, delete, match with core/text)	8	100.0
	Method of teaching (review, order of content)	6	75.0
	Use of manipulatives or technology	3	37.5
If no, why not?	No time	1	25.0
	Confident with current results	2	50.0
5. Have you made adjustments to your instruction of the pre-algebra curriculum based on previous years' results?	Yes	6	50.0
	No	6	50.0
If yes, what were the adjustments?	Content (add, delete, match with core/text)	3	50.0
	Method of teaching (review, order of content)	3	50.0
If no, why not?	Never previously received results	3	50.0
	Confident with previous results	2	33.3
	Apathy	1	16.7
	Don't know	1	25.0

When asked about previous years, half the teachers indicated that they had made adjustments. The adjustments were categorized as content and method alterations, similar

to those described for the current year. Of the teachers responding that previous adjustments had not been made, 50% said they did not make adjustments because they had not received results in prior years, 33.3% said they were satisfied with test results and did not think adjustments were warranted, 16.7% expressed apathy about test results. Teacher apathy is illustrated in the following quote:

> I'm not required to made adjustments, so I don't. I just cut off the grades. The test is a three-day filler at the end of the year. The kids think it's a joke and it is. It gives you something to do at the end. I guess that's a bad attitude, but I don't care how they do, and they don't care at all. They just fill in the blanks. The kids who care do fine, the kids who haven't cared all year just make connect the dot pictures. Why should I care?

More teachers reported using the UCAP results during the current year than in previous years. This may be due to the increased pressure felt by teachers to prepare students for the UCAP due to anticipated changes in its use with the implementation of the U-PASS accountability program.

Preparing Students for the Test

Teachers were asked three questions about the pressure to prepare students for the UCAP. Question 14 asked about the pressure to prepare students applied from parents, peers, and/or the principal of the school. Most teachers (66.7%) felt pressure to prepare students, yet one fourth of these said the pressure was self-imposed. The others emphasized the administrator's interest in student performance. One teacher responded that the pressure was from "my principal especially, and now the state." No teacher reported pressure from parents. The pressure applied by the principals is ironic considering that many teachers were not given results by the administration, were only invited to view

them, or had to find the results themselves, thereby making instructional adjustments for increased student achievement unlikely.

Questions 18 and 19 asked about the presence of consequences for students based on test performance, and the pressure, if any, this placed on teachers to prepare students for the test. Three quarters of the teachers disagreed that there were no student consequences since they applied the test score to the student's fourth term grade. Teachers using test scores as part of the student grade scored the tests at the school, a practice allowed by USOE. One third of the teachers who disagreed, however, felt that poor performance would have negative consequences on the students due to placement in future math courses. However, few teachers reported examining results of individual students, and only one mentioned placement of students as a use of results.

With the consequences in place, teachers felt pressure to prepare students to perform well on the UCAP. Only 16.7% of teachers disagreed, indicating that they do not feel any pressure to prepare students; 75.0% of the teachers indicated that they felt pressure to prepare students regardless of the presence or absence of consequences. One teacher who strongly agreed said, "I'm a professional! Of course I worry about it. Even with no consequences, which there are, adjustments are made. I really worry about the struggling kids. I want them to feel good about what they have learned." These results indicate that teachers build the UCAP into the pre-algebra course, including it as part of the course grade. This applies pressure to students and teachers alike. Teachers expressed an obligation to prepare students, even if consequences were not in place (see Table 35).

Table 35

Teacher Response to Question Concerning Preparing Students for the UCAP Test

Question	Response	n	%
14. I feel pressure from parents/peers/my principal to prepare my students to do well on the UCAP.	Strongly agree	2	16.7
	Agree	6	50.0
	Disagree	3	25.0
	Strongly disagree	1	8.3
	Not applicable /Don't know	0	0.0
18. The test results for the UCAP test have no consequences for the student.	Strongly agree	0	0.0
	Agree	3	25.0
	Disagree	4	33.3
	Strongly disagree	5	41.7
	Not applicable /Don't know	0	0.0
19. With or without consequences, I worry about adjusting my instruction for the benefit of my students' scores.	Strongly agree	1	8.3
	Agree	9	75.0
	Disagree	2	16.7
	Strongly disagree	0	0.0
	Not applicable /Don't know	0	0.0

Confidence in the UCAP and Instruction

Teachers were asked three questions about their confidence (a) in their current instruction including the type of items they use in class, and (b) that adjusting instruction would improve student test scores. To determine the overall opinion of the UCAP, teachers were also asked about their students' performance on the UCAP, and if they felt it was a valid indication of the students' ability (see Table 36).

Teachers felt confident that their current instruction was sufficient. Of the nine teachers who felt confident, two of them (22.5%) emphasized that there was "always room for improvement." Two teachers, who lacked confidence in their current instruction,

Table 36

Teacher Responses to Questions Concerning Confidence in UCAP and Instruction

Question	Response	n	%
15. The items on the UCAP are similar to the type of problems my students see during the year.	Strongly agree	1	8.3
	Agree	9	75.0
	Disagree	2	16.7
	Strongly disagree	0	0.0
	Not applicable /Don't know	0	0.0
16. My instruction of the pre-algebra curriculum is sufficient as it is currently implemented.	Strongly agree	2	16.7
	Agree	7	58.3
	Disagree	2	16.7
	Strongly disagree	0	0.0
	Not applicable /Don't know	1	8.3
17. I have confidence that making adjustments to my instruction will result in higher test scores.	Strongly agree	1	8.3
	Agree	8	66.7
	Disagree	2	16.7
	Strongly disagree	1	8.3
	Not applicable /Don't know	0	0.0
10. I have confidence that the results of the UCAP are a valid indication of my students' ability in pre-algebra.	Strongly agree	1	8.3
	Agree	4	33.3
	Disagree	4	33.3
	Strongly disagree	2	16.7
	Not applicable/Don't know	1	8.3
9. My students performed well on the UCAP.	Strongly agree	2	16.7
	Agree	8	66.7
	Disagree	1	8.3
	Strongly disagree	0	0.0
	Not applicable/Don't know	1	8.3

Note. Percent is based on number responding to question. UCAP = Utah Core Assessment Pre-Algebra.

cited the lack of time to cover the necessary content. "I never have enough time. I follow the core, and it's good, but not perfect. It takes more time than I have."

When asked about items found on the test and those used in class, 83.3% agreed

or strongly agreed that there was a match. However, 40% of these teachers also qualified their agreement with comments that item content matched well, but not item format. They felt that the content of the questions was similar, but the format used on the test was different from what was presented in class. Several cited alignment of item content and format as a recent adjustment made based on student test scores. One teacher, however, felt worried about this type of adjustment, stating:

> I worry about teaching to the test. I make a special lesson to show them how the format is and how the wording will be. They don't always find the right answer based on the calculation, then they just guess. Even if they could do the problems in class, the wording sometimes throws them off.

After expressing their level of confidence in current instruction, teachers were asked if they were confident that adjustments would lead to higher test scores. The majority (75.0%) agreed. However, there were several qualifications made to their response. Teachers felt that there were many other factors that influence test scores, including motivation, parental help, and student completion of assignments and homework. The teachers who disagreed cited many of the same factors, but felt that instruction could not overcome these obstacles to good student test performance. Another teacher raised the issue of teaching to the test, disagreeing with the notion of adjusting instruction altogether, "But then it [adjusting instruction] is a matter of teaching to the test. I don't agree with teaching to the test, so I don't adjust what I do to match it. That's not right."

Finally, teachers were asked if they felt their students had performed well on the UCAP. All but one teacher agreed that their students had performed well. Teachers described student performance as "better than I thought they would [perform]," "they

know even more than what's on the test," and "it's good, especially compared to last year." This confidence, however, may be based on teachers' interpretation of performance based on a comparison to other students in the district or state, as previously described.

When asked if the test was a valid indication of student ability, however, teachers were less enthusiastic. Only 41.7 % agreed, while 50.0 % disagreed, and 8.3 % were undecided. Those that disagreed were concerned that decisions about students based on one test could not be valid. "It only covers part of what they know, not everything," one teacher worried. "It's given under stress at the end of the year. That's why I don't think it is [a valid indication of ability]." Another teacher said, "One test can't decide if the kids really know it. The learning styles of the students are not considered, so it can't be the best way to decide if they know math." Other concerns dealt with the emphasis of certain problems on the test:

> It's not weighted. What is stressed on the test isn't balanced. The equation and properties are not covered the way they are in the book. There might be four questions on equations and two on properties, but the properties aren't even covered for one chapter and equations are half the course. It doesn't really measure ability, just if they can take a test.

Another expressed a similar concern:

> The time spent on some parts of the test don't always match the class. There are a lot of questions on graphing, but that is only one chapter in the book. And functions, graphing, proportions and statistics, maybe one chapter on each, but they are really emphasized on the test.

While teachers expressed frustration with the mismatch of UCAP items to the textbook coverage of the content, it should be noted that the UCAP items were intended to match the Utah Core Curriculum, not a specific textbook.

CHAPTER V

DISCUSSION, CONCLUSIONS, AND IMPLICATIONS

This study sought evidence to support assumptions underlying the purpose and use of the UCAP. To collect evidence, test content was examined including item match to course objectives, reliability, and subtest intercorrelations. Next, analyses of correlations of the UCAP with convergent and discriminant measures were completed, including an examination of both the pattern of correlations and tests of statistical significance. Finally, teachers were interviewed to ascertain the degree to which UCAP results were used to make necessary adjustments to instruction.

Discussion of Evidence

A discussion of the results is presented and synthesized here for each of the two arguments that guided this study: (a) UCAP test scores indicate seventh-grade students' level of mastery of knowledge and skill in pre-algebra; and (b) pre-algebra teachers use UCAP results to adjust instruction, leading to higher student achievement. Table 37 summarizes the results.

Argument One: Do UCAP Scores Indicate Mastery of Pre-Algebra?

The unified concept of validity guides collection of evidence based on the purpose

Table 37

Summary of Results by Arguments and Assumptions

Argument	Assumptions	Results
UCAP test scores indicate 7th-grade students' level of mastery of knowledge and skill in pre-algebra.	Assumption 1. UCAP content is relevant to and representative of the Utah Core Curriculum for Pre-algebra.	• 65% of Utah Core Curriculum objectives measured. • 49 of 80 items assigned to only 10 objectives. • Reliability estimates were high and consistent with other measures of mathematics.
	Assumption 2. Answering UCAP items correctly requires knowledge and skills of pre-algebra mathematics and is therefore considered a measure of pre-algebra.	• Strong positive correlations with convergent measures. • Convergent correlations were statistically significantly higher than discriminant measures.
Pre-algebra teachers use UCAP results to adjust instruction, leading to higher student achievement.	Assumption 4. Results are provided to teachers in a timely and meaningful report.	• Reports, which provided meaningful information, were returned to teachers just prior to the beginning of school.
	Assumption 5. UCAP scores are properly interpreted by teachers	• Teachers interpreted UCAP scores in a norm-referenced manner.
	Assumption 6. Teachers make appropriate and meaningful instructional adjustments based on UCAP scores.	• Instructional adjustments were made by some teachers, with review of content cited most often.

and use of test scores. Methods for collecting such evidence are well known. Unfortunately, as reported in technical manuals, only sparse evidence has been collected for tests currently being used by states for accountability purposes. Data have been collected based on specific validity types, for example, content or construct. Content

evidence provided by all states included item development and internal consistency reliability estimates.

A well-established method of collecting evidence that a test measures what it purports to measure is the correlation of test scores with both convergent and discriminant measures. The use of convergent and discriminant correlational evidence was supported in literature concerning collection of validity evidence, yet the six reviewed states had not collected adequate evidence to support the use of state tests as measures of mathematics knowledge and skill. Of the eight tests reviewed, five (62.5%) provided correlational evidence to support the assumption that the test measured what it purported to measure. Correlation coefficients were not analyzed or interpreted. Three of these tests relied on a single correlation coefficient as evidence: the two Texas tests and the Virginia test. The UCAP was among three state tests that provided no correlational evidence to support the assumption that it was a measure of pre-algebra. None of the technical manuals reported correlations for discriminant measures.

Similar to the six reviewed states, Utah presented the item development process and reliability as evidence for the validity of UCAP scores as indicators of student knowledge and skill in pre-algebra. Although these lines of evidence are important, they are insufficient evidence to support the use of the UCAP as a measure of student mastery of pre-algebra. As described in the evaluative argument for this study, item development and estimates of reliability address only a single assumption; other lines of evidence such as correlational data are needed.

This study examined the match of UCAP items to the Utah Core Curriculum for pre-algebra that indicated poor (65%) coverage of course objectives. The relative importance of objectives was neither reflected in the number of objectives for each subtest nor the number of objectives measured for each subtest. The lack of complete coverage of core objectives calls into question the usefulness of the UCAP as a measure of students' mastery of pre-algebra as described by the Utah Core Curriculum.

The WestEd evaluation of the UCAP, described in the review of literature, concluded that reliability of the total test was reasonable. However, WestEd did not analyze reliability values for the content and skill subtests nor were coefficients adjusted to allow for direct comparison. Use of the Spearman Brown prophecy formula to adjust subtest reliability coefficient values allowed for this comparison. Each of the content and skill subtests had similar reliability coefficients when adjusted for differing test lengths. The subtest coefficients were also comparable to adjusted SAT-9 math subtest coefficients, and to those of other state tests as described in the review of literature.

While there was some evidence of technical quality, the UCAP lacks sufficient items to have confidence that it is a measure of students' knowledge and skill in pre-algebra, or provide adequate information to teachers about instruction of the breadth of the course. Inferences made about students or instruction based on these test scores are likely to be erroneous, irrelevant, or detrimental due to the lack of complete representation of course content.

Pairwise comparisons of convergent correlation coefficients revealed that the expected pattern was found for the sample, but not for all subgroups. The standardized

mean differences for the UCAP and SAT-9 math subtest were similar to one another for each subgroup and the sample. The course grade standardized mean difference, however, was dissimilar to both math tests. This may indicate that student performance that led to the assigning of course grades was not reflective of performance on the two math tests. For example, for the subgroup based on reading proficiency, the standardized mean difference for grades was 1.37, less than the standardized mean differences of 2.03 and 2.27 for the UCAP and SAT-9 math subtest, respectively. This may indicate that the two math test scores are impacted to a greater degree by the reading proficiency of students than the course grade.

The analysis discussed thus far was important, but could not provide evidence that there was a distinction between the convergent and discriminant correlations. Two inferential tests were used to examine the relationship of convergent and discriminant measures. The null hypothesis for the test of homogeneity was rejected for all groups, indicating that statistically significant differences were present for the convergent and discriminant measures. This allowed for the test of discrimination to be completed. Rejection of the null hypothesis of this test indicated that, for the sample and each subgroup, the set of convergent correlation coefficients were statistically significantly higher than the set of discriminant coefficients. This provides support for the assumption that the UCAP is a measure of mathematics.

The analysis of correlation coefficients using both pattern examination and inferential tests moves beyond use of a single correlation to support the UCAP as a measure of mathematics. Tests of statistical significance provide strong support for the

assumption that UCAP measures some important components of pre-algebra. However, two aspects of the analysis diminish the strength of the argument that UCAP is a measure of pre-algebra knowledge and skill. First, lack of items for objectives within the core curriculum and use of single items to measure many of the objectives limit the inferences that can be made about students' abilities. Second, standardized mean differences in test scores for UCAP and SAT-9 math subtest suggest that, for this study sample, performance on these two measures provided similar information about students' pre-algebra knowledge and skill. It is not known if the results would be similar for students with different demographic characteristics, from mixed grade classes, or from those who had been taught using different textbooks and instructional materials. In addition, the ability of the SAT-9 math subtest to provide specific information pertaining to pre-algebra content as defined by the Utah core curriculum is limited.

Argument Two: Do Teachers Use UCAP Scores To Make Instructional Decisions?

The review of research showed an alarming paucity of published research or state reported collection of evidence concerning teacher use of standardized achievement test scores for making instructional decisions. The seven reviewed articles revealed that teachers used test scores only to a small degree. Authors cited teachers' inability to interpret results and lack of confidence in the tests ability to adequately measure student learning as possible reasons for the lack of use.

Teachers in this study did not feel that the report was returned in a timely manner.

Most often teachers received the report just prior to the start of the subsequent school year or after it began. They felt it was not possible to review the results and make adjustments to the curriculum. This finding was similar to findings of Yakimowski (1996) and Salmon-Cox (1981). Both authors found that late return of results hindered teachers' use of test scores, but it was not the primary reason for lack of use.

The UCAP score report provided information that teachers found helpful in determining the performance of students. The teachers' ability to interpret test scores was not well supported. In this study 91.6% of teachers interpreted performance of students based on comparison to students at other schools rather than by the percentage correct. Teachers' knowledge of measurement was not assessed through this interview so it is unclear if this presented an obstacle for teachers interpretation of test scores. Lack of adequate knowledge was an issue in the work reviewed by Green and Williams (1989), Marso and Pigge (1992), and Yakimowski (1996).

In spite of teachers receiving and interpreting student scores, adjustments made to instruction were found to be weak and meaningless. Similar to findings of Wilson and Corbett (1991) in which teachers cited review of content as the primary instructional adjustment, aimed at raising scores, not necessarily improving student understanding, teachers in this study had made adjustments to reviewing content or making simple adjustments to content order. Those teachers who reported instructional changes concerning content, use of instructional technology, or pedagogy did not provide adequate or convincing details about the changes, leading to a lack of sufficient evidence to support the assumption concerning teacher use of UCAP scores. Three possible explanations for

the lack of substantive changes are (a) lack of confidence in the test as a valid measure of student achievement, (b) lack of confidence that making instructional changes will lead to higher test scores, or (c) that teachers felt confident in the current performance of students and in their instruction as indicated by 83.4% and 75.0% of teachers, respectively. Similar to Nolen et al. (1992), Wilson and Corbett (1991), and Yakimowski (1996), this study found that 50.0% lacked confidence in the UCAP as a valid measure and therefore did not feel instructional adjustments were necessary. One quarter of teachers also felt that adjusting instruction would not necessarily increase student achievement, citing student motivation and completion of homework as factors that had greater influence on student achievement than instruction. Unfortunately, teachers did not view these factors being influenced by instruction. Finally, teachers did not cite instructional strategies that differed from those used in the past, only that order of introducing content or emphasis on content was changed.

Limitations

The UCAP has been in use for over 10 years and very little data had been collected to support its use. This study collected important evidence concerning validity of UCAP scores for making decisions about student knowledge of pre-algebra or instruction. However, validity is an ongoing process that requires new evidence when the use or purpose of a test changes. This research provided (a) an evaluative argument framework for collecting evidence, and (b) initial evidence pertaining to underlying assumptions

concerning the UCAP. These should be used to continue collecting evidence for tests of student achievement.

Evidence collected in this study adds substantially to information known about UCAP and its use by teachers. This information is important for understanding the impact of student testing as part of the existing accountability program in Utah, and has implications for implementation of U-PASS. To the extent that students and teachers in the participating district are systematically different from other students, results may not be generalizable to the state. For example, schools or districts that have different demographics, mixed grade classes, offer pre-algebra in grades other than seventh grade, or use different textbooks and instructional materials may have different results. Expanding the study to the entire state of Utah or using a representative sample would strengthen the conclusions of this study. Expansion of data collection to include teachers from throughout the state of Utah would provide more accurate information pertaining to teacher use of UCAP scores.

Conclusions and Implications

Based on results of this study, it was concluded that: (a) some evidence exists that the UCAP is a good measure of a limited number of pre-algebra course objectives as defined in the Utah core curriculum, and (b) teacher interpretation and use of UCAP scores for the purpose of making instructional decisions was limited and instructional adjustments were not meaningful.

Given that the UCAP is not a complete measure of the pre-algebra course objectives, the usefulness of scores for making instructional decisions is limited to the measured objectives. The UCAP does not appear to provide substantially different information than the SAT-9 math subtest for making decisions about student mastery of pre-algebra knowledge and skills. Furthermore, teachers in this study reported interpreting the UCAP in a normative manner, comparing their students' performance to others rather than a standard of performance. Teachers also reported making only general adjustments to instruction based on test scores. These general adjustments did not take advantage of the specific information provided by the UCAP report. Although this may call into question the use of both tests, the proper interpretation and use of UCAP test scores may lead to important instructional adjustments.

Conclusions of this study were based on use of UCAP as of the 1999-2000 school year, prior to the implementation of U-PASS, which mandates the public reporting of UCAP. Implementation of U-PASS represents an increase in the stakes for UCAP. Based on the review of literature, it is likely that teachers will make more instructional adjustments as the stakes of UCAP increase. This study found there was little evidence that teachers interpreted the UCAP score report completely or made meaningful instructional adjustments. If more adjustments are made due to increased pressure to raise test scores, but adjustments are not appropriate, it is doubtful that there will be any positive impact on student knowledge or skill in pre-algebra.

Recommendations for Further Research

The results of this study lead to several important issues that should be part of further investigation. First, teachers reported that the proportion of items for each subtest was not reflective of the importance of topics in the pre-algebra course. Similarly, review of item-objective match revealed that the number of objectives for each subtest was approximately equal, with no indication of relative importance. The current structure of the core curriculum needs revisions to accurately reflect the relative importance of concepts. Subsequently, UCAP's construction and item allotment should be reviewed and adjusted to more accurately reflect the important aspects of key course objectives.

Additional evidence is needed for both arguments described in this study. Data collection should be expanded to include a representative sample of students and teachers in Utah. Additional methods described in the review of literature, but not included in this study should also be applied. In particular, item analysis should be extended to include process analysis to further investigate the assignment of items to specific subtests. Information gained during this process would provide valuable information to teachers about the cognitive processes involved in answering UCAP items. Evidence of teacher use should include the investigation of the link between instructional methods and student achievement, the ability of teachers to interpret test scores, and the subsequent design and implementation of adjustments needed for instruction.

Richness of information available through observation and investigation of teacher behavior in the classroom would add significantly to evidence of test use. These data could

also be used to investigate the factors that influence the use of test scores such as professional development opportunities, significance placed on testing by the principal and teachers, and the content and pedagogical knowledge of teachers.

As validity is an ongoing process, based on the purpose and use of a test, the further research recommended is vital to support the continued use of the UCAP as part of the U-PASS accountability system.

Final Comments

In the current climate of educational accountability through testing, it is safe to assume that the UCAP, or other similar tests, will become increasingly important in making decisions about student knowledge and teacher effectiveness. Investigations of the appropriateness of inferences about students and test use must be on-going components of a quality testing program.

REFERENCES

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. Psychological Bulletin, 51(2, Pt. 2), 2.2.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1974). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Anastasi, A. (1982). Psychological testing (5th ed.). New York: Macmillan.

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. Braun (Eds.), Test validity (pp. 19-32). Hillsdale, NJ: Erlbaum.

Baker. E. L. (1988). Mandated tests: Reform or quality indicator? (ERIC Document Reproduction Service No. ED 341 733). Los Angeles: University of California Graduate School of Education, CSE Dissemination Office.

Barton, P. E. (1999). Too much testing of the wrong kind; Too little of the right kind in K-12 education. Princeton, NJ: Educational Testing Service.

Berk, R. A. (1988). Fifty reasons why student achievement gain does not mean teacher effectiveness. Journal of Personnel Evaluation in Education, 1(4), 345-364.

Black, P. & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. Phi Delta Kappan [On-line.]. Available: http:// www.pdkintl.org/ kappan/kbla9810.htm

Bond, L. A. (1995). Norm-referenced testing and criterion-referenced testing: The differences in purpose, content, and interpretation of results. Oak Brook, IL: North Central Regional Educational Laboratory. (ERIC Document Reproduction Service ED 402 327)

California Department of Education. (2000). Alignment, validity, and reliability of the Spring 2000 Golden State Examination: A report to the senate, assembly, department of finance, state board of education. [On-line.]. Available: http://www.cde.ca.gov/statetests/gse/gsereliabilityrpt.pdf

Canady, R. L., & Hotchkiss, P. R. (1989). It's a good score! Just a bad grade. Phi Delta Kappan, 71, 68-71

Council of Chief State School Officers. (1998). Key state education policies on K-12 education. Washington, DC: Author.

Crocker, L., & Algina, J. (1986). Introduction to classical & modern test theory. Orlando, FL: Harcourt Brace Jovanovich.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), Educational measurement (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L. J. (1988). Five perspectives on validation argument. In H. Wainer & H. Braun (Eds.), Test validity (pp. 3-17). Hillsdale, NJ: Erlbaum.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), Intelligence: Measurement theory and public policy: Proceedings of a symposium in honor of Lloyd G. Humphreys (pp. 147-171). Urbana: University of Illinois Press.

Cronbach, L. J., & Meehl, P.E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, 281-302.

Daniel, L.G., & King, D.A. (1998). Knowledge and use of testing and measurement literacy of elementary and secondary teachers. The Journal of Educational Research, 91, 331-344.

Etsey, Y. K. (1997, March). Teachers' and school administrators' perspectives and use of standardized achievement tests: A review of published research. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). Educational research: An introduction (6th ed.). White Plains, NY: Longman.

Geisinger, K. (1992). The metamorphosis of test validation. Educational Psychologist, 27, 197-222.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: some questions. American Psychologist, 18, 519-521.

Glaser, R., & Nitko, A. J. (1971). Measurement in learning and instruction. In R. L. Thorndike (Ed.), Educational measurement (2nd ed., pp. 625-670). Washington, DC: American Council on Education.

Green, K. E. (1992). Differing opinions on testing between preservice and inservice teachers. Journal of Educational Research, 86(1), 37-42.

Green, K. E., & Stager, S. F. (1985, April). Teachers attitudes toward testing. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Green, K.E., & Stager, S. F. (1986, April). Effects of training, grade level, and subject taught on the types of tests and test items used by teachers. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Green, K. E., & Williams, E. J. (1989, March). Standardized test use by classroom teachers: Effects of training and grade level taught. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Haertel, E. (1985). Construct validity and criterion-referenced testing. Review of Educational Research, 55, 23-46.

Haertel, E. (1986). The valid use of student performance measures for teacher evaluation. Educational Evaluation and Policy Analysis, 8(1), 45-60.

Haertel, E. (1999). Validity arguments for high-stakes testing: In search of the evidence. Educational Measurement: Issues and Practices, 18(4), 5-9.

Haladyna, T. M., Haas, N. S., & Allison, J. (1998). Continuing tensions in standardized testing. Childhood Education, 74, 262-273.

Haladyna, T.M., Haas, N.S., & Nolen, S.B. (1989). Tests score pollution (Technical Report 89-1). Phoenix: Arizona State University West.

Hambleton, R. K. (1980). Test score validity and standard-setting methods. In R. A. Berk (Ed.), Criterion-referenced measurement: The state of the art (pp. 80-123). Baltimore: Johns Hopkins University Press.

Hambleton, R. K. (1981). Contributions to criterion-referenced testing technology: An introduction. Applied Psychological Measurement, 4, 421-424.

Hambleton, R. K., & Rogers, H. J. (1991). Advances in criterion-referenced measurement. In R. K. Hambleton and J. N. Zaal (Eds.), Advances in educational and psychological testing (pp. 3-38). Boston: Kluwer Academic Publishers.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 48, 1-47.

Harcourt Brace Educational Measurement. (1997a). Stanford achievement test series, ninth edition: Spring norms book. San Antonio, TX: Harcourt Brace.

Harcourt Brace Educational Measurement. (1997b). Stanford achievement test series, ninth edition: Technical data report. San Antonio, TX: Harcourt Brace.

Hills, J. R. (1991). Apathy concerning grading and testing. Phi Delta Kappan, 72, 540-545.

Institute for Behavioral Research in Creativity (1999). Technical manual for the Utah State Office of Education core assessment series: Second edition, form A. Salt Lake City, UT: Author.

Kane, M. T. (1992). An argument-based approach to validity. Psychological Bulletin, 112, 527-535

Kentucky Department of Education. (1997). KIRIS accountability cycle 2 technical manual. Frankfort: Author.

Linn, R. L. (1980). Issues of validity for criterion-referenced measures. Applied Psychological Measurement, 4, 547-561.

Madaus, G. F. (1987). Testing and the curriculum. Chestnut Hill, MA: Boston College.

Madaus, G. F., & Tan, A. G. (1993). The growth of assessment. In G. Cawelti (Ed.), Challenges and achievements of American education (pp. 53-79). Alexandria, VA: Association for Supervision and Curriculum Development. (ERIC Document Reproduction Service ED 353 261)

Marso, R. N., & Pigge, F. L. (1992, February). Classroom teachers' perceptions of the extent and effectiveness of their schools' uses of standardized test results. Paper presented at the annual meeting of the Association of Teacher Educators, Orlando, FL.

McMillan, J. H., Myran, S., & Workman, D. (1999, April). The impact of mandated statewide testing on teachers' classroom assessment and instructional practices. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Meng, X. L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. Psychological Bulletin, 111, 172-175.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 30, 955-966.

Messick, S. (1980). Test validity and the ethics of assessment. American Psychologist, 35, 1012-1027.

Messick, S. (1984). The psychology of educational measurement. Journal of Educational Measurement, 21, 215-237.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. Educational Researcher, 18(2), 5-11.

Messick, S. (1993). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13-106). Phoenix: Oryx Press.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from person's responses and performances as scientific inquiry into score meaning. American Psychologist, 50, 741-749.

National Commission on Excellence in Education. (1983). A nation at risk: The imperative for educational reform. Washington, DC: U. S. Department of Education.

Nolen, S. B., Haladyna, T. M., & Haas, N. S. (1992). Uses and abuses of achievement test scores. Educational Measurement: Issues and Practice, 11(2), 9-15.

North Carolina Department of Public Instruction. (1996a). North Carolina end-of-grade tests: Reading comprehension, mathematics. Technical report no. 1. Raleigh, NC: Author. (ERIC Document Reproduction Service ED 406 397)

North Carolina Department of Public Instruction. (1996b). North Carolina end-of-course tests: Algebra I; Biology; Economic, legal, and political systems; English I; U.S. history (Technical report no. 1). Raleigh, NC: Author. (ERIC Document Reproduction Service ED 406 392)

Odden, A. (1986). Sources of funding for educational reform. Phi Delta Kappan, 67, 335-40.

Pedhazur, E. J., & Schmelkin, L. P. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Erlbaum.

Popham, W. J. (1978). Criterion-referenced measurement. Englewood Cliffs, NJ: Prentice-Hall.

Popham, W. J., & Husek, T. R. (1969). Implications of criterion-referenced measurement. Journal of Educational Measurement, 6, 1-9.

Salmon-Cox, L. (1981). Teachers and standardized tests: What's really happening? Phi Delta Kappan, 62, 631-634.

Shepard, L. A. (1993). Evaluating test validity. Review of Research in Education 19, 405-450.

Stevens, J. (1996). Applied multivariate statistics for the social sciences (3rd ed.). Hillsdale, NJ: Erlbaum.

Texas Education Agency. (1999). 1997-1998 technical digest [On-line.]. Available: http://www.tea.state.tx.us/student.assessment/techdig.htm

Utah State Office of Education. (1997). Core assessment handbook and specific directions for administration: Pre-algebra. Salt Lake City, UT: Author.

Utah State Office of Education. (1999). Fall enrollment by race/ethnicity, 1998-1999 school year, total all students.[On-line.]. Available: http://www.usoe.k12.ut.us/data/enrollment/ferace98.xls

Utah State Office of Education. (2000). Questions and answers about the Utah Core Assessment program [On-line.]. Available: http://www.usoe.k12.ut.us/eval

Watson, L. E. (1990). Educator perceptions of standardized achievement test uses and practices in the public schools of Idaho. Unpublished doctoral dissertation, University of Idaho, Moscow.

WestEd. (1999). The evaluation of the Utah state core curriculum and state core assessment documents in reading/language arts and mathematics. San Francisco, CA: Author.

Western Michigan University. (1995). An independent evaluation of the Kentucky Instructional Results Information System (KIRIS). Frankfort: Kentucky Institute for Education Research.

Wilson, B. L., & Corbett, H. D. (1991). Two state minimum competency testing programs and their effects on curriculum and instruction. Philadelphia, PA: Research for Better Schools. (ERIC Document Reproduction Service No. 377 251)

Yakimowski, M. (1996, April). Impact of state and federal student assessment legislation on curriculum, instruction, and student learning: The perspectives from California, Colorado, Connecticut and Illinois school districts. Paper presented at the annual meeting of the American Educational Research Association, New York.

APPENDICES

Appendix A:

Secondary Mathematics CRT Use by State and Grade Level of Administration

State	Grade Level/Subject
Alabama	9, 11, end-of-course algebra and geometry
Arkansas	11
California	end-of-course algebra and geometry
Connecticut	8, 10
Florida	11
Georgia	11, 12
Hawaii	credit by examination algebra
Kansas	7, 10
Kentucky	8, 11
Louisiana	7, 10
Maine	8, 11
Maryland	7 - 12
Michigan	7, 11
Minnesota	8 - 12
Missouri	8, 10
Mississippi	end-of-course algebra
New Hampshire	10
New Jersey	11, 12
New Mexico	10
New York	9 - 11
North Carolina	8 - 11

State	Grade Level/Subject
Ohio	8 - 12
Oklahoma	8, 11
Oregon	8, 10
Pennsylvania	8, 11
South Carolina	8, 10
Tennessee	6 - 8
Texas	6 - 8, 10 - 12
Utah	6 - 12
Vermont	8, 10
Virginia	8, 11
West Virginia	6 - 11

Note. Adapted from CCSSO, 1998, p. 20-21.

Appendix B:

Criteria for Determining Quality of Test Use Studies

Studies	Criteria for judging quality			
	A	B	C	D
Salmon-Cox (1981)	NS	3	1	1
Green and Williams (1989)	NS	3	NS	2
Wilson and Corbett (1991)	1	1	1	1
Marso and Pigge (1992)	NS	1	NS	2
Nolen, Haladyna, and Haas (1992)	1	2	NA	1
Yakimowski (1996)	NS	1	NS	2
McMillan, Myran, and Workman (1999)	1	1	1	1

Note. 1 = Excellent, 2 = Good, 3 = Poor, NS(4) = not specified, NA(0) = not appropriate due to confidentiality; low score indicated higher quality.

A. Instrument Field Tested/Piloted:

Development of Instrument reflected careful consideration of participants including field testing or piloting to make appropriate revisions.

B. Data Analyzed Thoroughly and Appropriately:

Survey and interview data reported using appropriate analysis. For example, if groups were compared, analysis should include appropriate analysis of group differences.

C. Member Check:

When appropriate, researcher should clarify participant answers with follow-up correspondence. This would also included non-response bias checks.

D. Conclusions Grounded in Data:

Authors' conclusions are well supported by data, including the sample size and response rate.

Appendix C:

UCAP Item Match to Utah Standards for Pre-Algebra

Standards and Objectives	Items
Number and Number Relationships	
1. Understand, represent, and use numbers in a variety of equivalent forms (integer, fraction, decimal, percent, exponential, and scientific notation) in real-world and mathematical problem situations.	32
2. Develop number sense for whole numbers, fractions, decimals, integers, and rational numbers.	55, 56
3. Understand and apply ratios, proportions, and percents in a wide variety of situations.	15, 16, 24, 25, 28, 55, 56
4. Investigate relationships among fractions, decimals, and percents.	13, 14
5. Represent numerical relationships in one- and two-dimensional graphs.	44, 45, 48, 49
Number systems and Number Theory	
1. Understand and appreciate the need for numbers beyond the whole numbers.	
2. Develop and use order relations for whole numbers, fractions, decimals, integers, and rational numbers.	11, 12
3. Extend their understanding of whole number operations to fractions, decimals, and integers; and rational numbers to scientific notation, exponents, and percents.	
4. Understand how the basic arithmetic operations are related to one another.	4, 18, 19, 20, 21, 22, 36
5. Develop and apply number theory concepts (e.g., primes, factors, and multiples) in real-world and mathematical problem situations.	1, 2, 5

Standards and Objectives	Items
Computation and Estimation	
1. Compute with whole numbers, fractions, decimals, integers, and rational numbers.	3, 8, 10, 11, 24, 25
2. Develop, analyze, and explain procedures for computation and techniques for estimation.	6
3. Develop, analyze, and explain methods for solving proportions.	
4. Select and use an appropriate method for computing from among mental arithmetic, estimation, paper-and-pencil, calculator, and computer methods.	17, 39
5. Solve problems by using computation, estimation, and proportionality.	7, 9, 26, 27, 29
6. Estimate to check the reasonableness of results.	
Patterns and Functions	
1. Describe, extend, analyze, and create a wide variety of patterns.	41, 42
2. Describe and represent relationships with tables, graphs, and rule.	46, 47, 50, 51, 53, 54
3. Analyze functional relationships to explain how a change in one quantity results in a change in another.	
4. Employ patterns and functions to represent and solve problems.	40, 49, 80
Algebra	
1. Understand the concepts of variable, expression, and equation.	30, 31
2. Represent situation and number patterns with tables, graphs, verbal, rules, and equations and explore the interrelationships of these representations.	32, 37
3. Analyze tables and graphs to identify properties and relationships.	
4. Develop confidence in solving linear equations using concrete and informal methods.	18, 19, 20, 22, 34, 38
5. Investigate inequalities and non-linear equations informally.	21, 23
6. Apply algebraic methods to solve a variety of real-world and mathematical problems.	51

Standards and Objectives	Items
Statistics	
1. Collect, organize, and describe data in a systematic fashion.	62, 63
2. Construct, read, and interpret tables, chars, and graphs.	52, 57, 58, 59, 60, 61
3. Make inferences and convincing arguments that are based on data analysis.	
4. Evaluate arguments that are based on data analysis.	
5. Develop an appreciation for statistical methods as a powerful means for decision making.	
Probability	
1. Model situations by devising and carrying out experiments or simulations to determine probabilities.	66, 67
2. Model situations by constructing a sample space to determine probabilities.	65
3. Compare experimental results with mathematical expectations in order to appreciate the power of using a probability model.	
4. Make predictions that are based on experimental or theoretical probabilities.	64, 68
5. Develop an appreciations for the pervasive use of probability in the real world.	
Geometry	
1. Identify, describe, compare, and classify geometric figures.	70, 71, 72, 73, 75
2. Visualize and represent geometric figures with special attention to developing spatial sense.	69, 74, 75, 76
3. Explore transformations of geometric figures.	
4. Represent and solve problems using geometric models.	
5. Understand and apply geometric properties and relationships.	29, 80

Standards and Objectives	Items
6. Develop an appreciations of geometry as a means of describing the physical world.	
Measurement	
1. Extend their understanding of the process of measurement.	
2. Estimate, make, and use measurements to describe and compare phenomena.	
3. Select appropriate units and tools to measure to the degree of accuracy required in a particular situation.	43
4. Understand the structure and use of systems of measurement.	
5. Extend their understanding of the concepts of perimeter, area, volume, angle measure, capacity, and weight and mass.	78, 79, 80
6. Develop the concepts of rates and other derived and indirect measurements.	27
7. Develop formulas and procedures for determining measures to solve problems.	77

Note. Adapted from USOE, 1997, pp. 18-21

Appendix D:

State UCAP Report Form

ASSESSMENT PROFILE

District/Area: STATE SUMMARY Test: PRE-ALGEBRA 2ND EDITION Grade Level: 07-08
School: Teacher: Test Number: 00000PA0708-N
Student Name: I.D.Number: Testing Date: SEPT. 30, 1999
Number Tested: 32,666

STANDARDS OR CONCEPTS	Student: Percent (%) Correct	Student: Minimum Expected % Correct	Student: Above/Below Expected %	Group Average: Class	Group Average: School	Group Average: Area	Group Average: District	Group Average: State	STUDENT PERCENT CORRECT COMPARED TO MINIMUM EXPECTED PERCENT CORRECT
									% Points Below Expected: < 45 | 30 | 15 | Minimum Expected % Correct: ↓ | % Points Above Expected: 15 | 30 | 45 >
NUMBER/NMBR RELATIONSHPS	52							52	
NUMBR SYSTMS/NMBR THEORY	53							53	
COMPUTATION/ESTIMATION	62							62	
PATTERNS AND FUNCTIONS	55							55	
ALGEBRA	61							61	
STATISTICS	69							69	
PROBABILITY	70							70	
GEOMETRY	64							64	
MEASUREMENT	50							50	
PROCEDURAL	58							58	
CONCEPTUAL	57							57	
APPLICATION	63							63	
TOTAL TEST	59							59	

Objectives	Number Correct/Possible	Percent Correct	Objectives	Number Correct/Possible	Percent Correct
NUMBER/NMBR RELATIONSHPS	7.8/15		CONCEPTUAL	18.8/33	
PROPORTION	3.6/6		APPLICATION	13.3/21	
PERCENT	2.2/4				
GRAPHS	1.9/5				
NUMBR SYSTMS/NMBR THEORY	6.4/12				
ORDER RELATIONS	.9/2				
NUMBER THEORY	1.7/4				
PROPERTIES OF EQUALITY	3.8/6				
COMPUTATION/ESTIMATION	8.7/14				
ORDER OF OPERATIONS	.8/1				
FRACTIONS, COMPUTE	4.0/6				
PRBLM SLVING PROPORTION	2.2/4				
ESTIMATION	1.7/3				
PATTERNS AND FUNCTIONS	5.5/10				
NUMBER PATTERS	2.1/4				
FUNCTION PATTERS	3.3/6				
ALGEBRA	8.5/14				
1 STEP EQATIONS/INEQUAL	2.6/4				
OTHR EQUATIONS/EQUALTIE	2.6/4				
UTILIZE MANIPULATIVES	1.1/2				
TRNSLATE VERBAL/ALGEBRA	2.3/4				
STATISTICS	4.8/7				
PROBABILITY	3.5/5				
GEOMETRY	5.8/9				
MEASUREMENT	2.5/5				
PROCEDURAL	15.0/26				

Appendix E:

Pre-Algebra Teacher Interview Protocol

Teacher ID_____
School ID_____
Gender M F
Years of Teaching Experience_____
Pre-Algebra sections/day (1998-99 year)_____
1. Did you receive the score reports for the May 1999 administration of the Utah Core Assessment Pre-Algebra test?
Yes No
1b. If yes, were the results reported by student?_____teacher_____ school?_____
Comments:

2. If received, did you carefully examine the results for each student?_____
for your own class(es)?_____for the school?_____
Comments:

3. If received, did your department, as a group, carefully examine the results for the school?
Yes No
3b. If yes, did the department discuss areas of concern? Yes No
3c. If yes, what were the areas of concern?
Comments:

4. Based on the results on the summary reports (either teacher or school), have you made adjustments, to your instruction of the pre-algebra curriculum?
Yes No

4b. If yes, what were the adjustments?

4c. If no, why not?

5. Have you made adjustments to your instruction of the pre-algebra curriculum based on previous years' results?
Yes No

5b. If yes, what were the adjustments?
5c. If no, why not?

6. Using this sample report form as an example, what information on the report is helpful to you in determining how well your students performed on the Utah Core Assessment Pre-Algebra Test?

7. Which information, the actual percent correct or the comparison to district and state percent correct, is most important to you in determining how well your students performed on the Utah Core Assessment Pre-Algebra Test?

8. Using this sample report form as an example, what information is helpful to you in determining what adjustments should be made to your pre-algebra instruction?

Respond to each of the following questions by indicating whether you:

Strongly agree	Agree	Disagree	Strongly disagree	Not Applicable/Don't Know
1	2	3	4	0

_____9. My students performed well on the Utah Core Assessment Pre-Algebra Test.

_____10. I have confidence that the results of the Utah Core Assessment Pre-AlgebraTtest are a valid indication of my students' ability in pre-algebra.

_____11. The results of the Utah Core Assessment Pre-Algebra Test are returned in a timely manner to allow for adjustments to the instruction of the pre-algebra curriculum to be made.

_____12. I have confidence in my ability to interpret the results of the Utah Core Assessment Pre-Algebra Test.

_____13. The information provided on the test report facilitates adjustment to my pre-algebra instruction.

_____14. I feel pressure from parents/peers/my principal to prepare my students to do well on the Utah Core Assessment Pre-Algebra Test.

_____15. The items on the Utah Core Assessment Pre-Algebra Test are similar to the type of problems my students see during the year.

_____16. My instruction of the pre-algebra curriculum is sufficient as it is currently implemented

_____17. I have confidence that making adjustments to my instruction would result in higher test scores.

_____18. The test results for the Utah Core Assessment Pre-Algebra Test have no consequences for the student.

_____19. With or without consequences, I worry about adjusting my instruction for the benefit of my students' scores.

Appendix F:

Correlation of Convergent and Discriminant Measures for the Sample and All Subgroups

Table F1

Convergent and Discriminant Evidence: Correlation of Scores on the UCAP, SAT-9 Subtests, Pre-Algebra Course Grade, and Teacher Rating of Pre-Algebra Knowledge for Sample

Measures	1	2	3	4
Convergent				
1. UCAP total score	--			
2. SAT-9 math total score	.73	--		
3. Course grade	.61	.53	--	
4. Teacher rating	.75	.70	.86	--
Discriminant (SAT-9 subtest)				
5. Reading	.56	.60	.38	.63
6. Language	.59	.65	.46	.66
7. Science	.54	.62	.33	.53
8. Social Science	.51	.54	.35	.56
9. Listening	.48	.49	.33	.49

Note. UCAP = Utah Core Assessment Pre-Algebra; SAT-9 = Stanford Achievement Test, 9th Edition.

Table F2

Convergent and Discriminant Evidence: Correlation of Scores on the UCAP, SAT-9 Subtests, Pre-Algebra Course Grade, and Teacher Rating of Pre-Algebra Knowledge for Males and Females

Measures	1		2		3		4	
Convergent								
1. UCAP Total score	--		.71		.62		.69	
2. SAT-9 math total score	.74		--		.54		.74	
3. Course grade	.62		.54		--		.83	
4. Teacher rating	.80		.69		.88		--	
Discriminant (SAT-9 subtests)	F	M	F	M	F	M	F	M
5. Reading	.58	.54	.62	.60	.67	.39	.61	.64
6. Language	.59	.60	.66	.67	.42	.47	.65	.64
7. Science	.55	.57	.62	.62	.36	.37	.56	.51
8. Social Science	.48	.55	.54	.53	.33	.39	.57	.56
9. Listening	.50	.45	.50	.48	.34	.32	.45	.51

Note. Correlations for male participants (n = 655) are presented above the diagonal, and correlations for females participants (n = 806) are presented below the diagonal.
UCAP = Utah Core Assessment Pre-Algebra; SAT-9 = Stanford Achievement Test, 9th Edition.

Table F3

Convergent and Discriminant Evidence: Correlation of Scores on the UCAP, SAT-9 Subtests, Pre-Algebra Course Grade, and Teacher Rating of Pre-Algebra Knowledge for Ethnic Minority and Majority Students

Measures	1		2		3		4	
Convergent								
1. UCAP total score	--		.66		.67		1.00[a]	
2. SAT-9 math total score	.73		--		.56		.77	
3. Course grade	.61		.53		--		1.00[a]	
4. Teacher rating	.75		.70		.86		--	
Discriminant (SAT-9 subtests)	Ma	Mi	Ma	Mi	Ma	Mi	Ma	Mi[a]
5. Reading	.56	.48	.60	.60	.38	.36	.63	.84
6. Language	.60	.50	.65	.63	.46	.45	.65	.95
7. Science	.55	.55	.61	.69	.33	.33	.52	.93
8. Social Science	.51	.58	.53	.62	.35	.40	.55	.80
9. Listening	.47	.44	.48	.55	.32	.36	.48	.62

Note. Correlations for minority students (n = 90) are presented to the right and above the diagonal, and correlations for majority students (n = 1,371) are presented to the left and below the diagonal. Mi = minority students; Ma = majority students; UCAP = Utah Core Assessment Pre-Algebra; SAT-9 = Stanford Achievement Test, 9th Edition.
[a]number of minority students rated = 3.

Table F4

Convergent and Discriminant Evidence: Correlation of Scores on the UCAP, SAT-9 Subtests, Pre-Algebra Course Grade, and Teacher Rating of Pre-Algebra Knowledge for Reading Proficiency Readers of Level 1 and 4

Measures	1		2		3		4	
Convergent								
1. UCAP total score	--		.65		.66		.[a]	
2. SAT-9 math total score	.53		--		.53		.[a]	
3. Course grade	.42		.29		--		.[a]	
4. Teacher rating	.38		.31		.58		--	
Discriminant (SAT-9 subtests)	L1	L4	L1	L4	L1	L4	L1	L4
5. Reading	.04	.27	-.02	.17	.21	.08	.30	.[a]
6. Language	.29	.32	.29	.36	.30	.29	.42	.[a]
7. Science	.31	.42	.34	.30	.18	.25	-.26	.[a]
8. Social Science	.28	.25	.30	.11	.21	.17	.08	.[a]
9. Listening	.28	.21	.28	.21	.21	.25	-.00	.[a]

Note. Correlations for reading proficiency level 4 (n = 75) are presented above the diagonal, and correlations for reading proficiency Level 1 (n = 113) are presented below the diagonal. L1 = Level 1 reading proficiency; L4 = Level 4 reading proficiency. UCAP = Utah Core Assessment Pre-Algebra; SAT-9 = Stanford Achievement Test, 9th Edition.

[a] Rating was constant (master) for all students with Level 4 reading proficiency.

Table F5

Convergent and Discriminant Evidence: Correlation of Scores on the UCAP, SAT-9 Subtests, Pre-Algebra Course Grade, and Teacher Rating of Pre-Algebra Knowledge for Masters and Nonmasters of Pre-algebra

Measures	1		2		3		4
Convergent							
1. UCAP total score	--		.71		.47		[a]
2. SAT-9 math total score	.52		--		.29		[a]
3. Course grade	.70		.47		--		[a]
4. Teacher rating	[a]		[a]		[a]		[a]
Discriminant (SAT-9 Subtests)	NM	M	NM	M	NM	M	
4. Reading	.33	.55	.44	.45	.25	.25	.[a]
5. Language	.47	.55	.58	.54	.36	.24	.[a]
6. Science	.47	.58	.52	.37	.39	.21	.[a]
7. Social Science	.51	.32	.54	.36	.36	.12	.[a]
8. Listening	.40	.39	.37	.35	.37	.22	.[a]

Note. Correlations for masters (n = 71) are presented above the diagonal, and correlations for nonmasters (n = 64) are presented below the diagonal. NM = nonmaster; M = master. UCAP = Utah Core Assessment Pre-Algebra; SAT-9 = Stanford Achievement Test, 9th Edition.

[a] Rating variable was constant, thus omitted.

VITA

Louise Richards Moulding

Education

2000	Ph.D., Research & Evaluation Methodology, Utah State University
1995	M.Ed., Curriculum & Instruction, Weber State University
1989	B.S., Biology composite major, Chemistry minor, Weber State University

Professional Experience

2000-present	**Curriculum Standards Team Leader**, Davis School District
1998-present	**Evaluation Consultant**, Biological Sciences Curriculum Study (BSCS), Colorado Springs, CO (integrated science curriculum design study)
1995-2000	**Curriculum Coordinator**, Secondary Mathematics, Science, Language Arts, Weber School District
1995-2000	**MESA Administrator** (Math, Engineering, Science Achievement program for historically under-represented students), Weber School District
1995-2000	**Eisenhower Professional Development Program Coordinator**, Weber School District
1994-2000	**Director**, Weber County School District Science Fair
1997-1998	**Co-Director**, Earth Systems Science Core Experiment Development, Utah State Office of Education.
1996-1999	**Coordinator**, Weber State University Teacher Academy, Mathematics, Science, and Language Arts
1994-1995	**Assessment Consultant**, Chemistry End-of-Course Test Development, Utah State Office of Education
1995-1996	**Instructor**, Science Performance Assessment, Brigham Young University
1991-92	**Assessment Consultant**, Human Biology End of Course Test Development and Test Item Pool Development, Institute for Behavioral Research in Creativity (IBRIC),Salt Lake City, UT
1989-1998	**Teacher**, Chemistry, Biology , Advanced Placement Biology, Grades 9-12, Ogden City and Weber County School Districts
1990-1992	**Instructor**, Summer Science Camp, Science Technology, and Society, Brigham Young University
1989-1990	**Instructional Materials Writer**, Science Technology and Society Textbook, Utah State Office of Education

Professional Presentations

Moulding, L. (2000, November). Utilization of state criterion-referenced test scores: Does testing lead to instructional adjustments? Paper presented at the American Evaluation Association Annual Convention, Honolulu, HA.

Moulding, L. & Drickey, N. (2000, April). Process and product evaluation: Enhancing instruction and assessment in the science classroom. National Science Teachers Association National Conference, Orlando, FL.

Moulding, L. (2000, January). Assessing classroom performance: Strategies for first year science teachers. First Year Science Teacher Workshop, Weber State University

Wareham, K., Moulding, L., & Puhlmann, N. (1999, November). Evaluation training programs: Is the foundation solid? Panel presentation at the American Evaluation Association Annual Convention, Orlando, FL.

Moulding, L. (1999, April). Evaluating an integrated science design study: Challenges and successes. American Educational Research Association Annual Conference, Montreal, Quebec, Canada.

Moulding, L. (1998, April). Aligning assessment to standards. National Science Teacher Association National Conference, Las Vegas, NV.

Moulding, L. (1995, April). Implementing the new secondary science core curriculum. National Science Teacher Association Regional Conference, Salt Lake City, UT.

Moulding, L. (1994, December).Reverse dot blot laboratory procedure for A.P. Biology students. Utah Science Teacher Workshop Midwinter Conference, Salt Lake City, UT.

Refereed Publication

Smith, J. A., Puhlmann, N., Jones, S.C., Moulding, L., Elwell, C., & Morgan, W. (1998). Implementing an arts education program at Edith Bowen Laboratory School: What we've learned after one year. National Association of Lab Schools, pp. 6-10.

Evaluation Reports

Moulding, L. (2000). Making sense of integrated science: A guide for high schools. A final report presented to Biological Science Curriculum Study. Logan, UT

Puhlmann, N., Elwell, C., Jones, S., & Moulding, L. (1998). Edith Bowen Laboratory School arts program: Year 1 evaluation results from interviews, focus groups, and surveys. Logan, UT

Awarded Grants

Electronic High School Human Biology Course Development, Technology Literacy Grant, (1998-1999); U.S. Department of Education and Utah State Office of Education, $12, 500.

Seventh and Eighth Grade Integrated Science "Sci-ber Text": Electronic Resources for Teachers, (1997-1999); Utah State Office of Education, $30,000.

Using Technology in the A.P. Environmental Classroom, (1997-1998); Utah State Office of Education, $8,000.

Ninth Grade Earth Systems Core Experiment Development, (1997-1998). Utah State Office of Education, $20,000.

Awards

1997-2000	**Fellow**, National Science Foundation (NSF)/ American Educational Research Association (AERA) Evaluation Training Program, Utah State University
1993-1997	**Harold W. Ritchie Award** for Outstanding Direction of Students in Science Fair, Weber State University

Professional Association Membership

American Educational Research Association
American Evaluation Association
American Psychological Association
Division Five: Evaluation, Measurement, and Statistics
National Science Teacher Association
National Council of Teachers of Mathematics
Utah Council of the International Reading Association
Utah Science Teachers Association